DOCUMENT RESUME

ED 417 199 TM 028 140

AUTHOR Bay, Luz; Chen, Lee; Hanson, Bradley A.; Happel, Jay; Kolen, Michael J.; Miller, Timothy; Pommerich, Mary; Sconing, James; Wang, Tianyou; Welch, Catherine

TITLE ACT's NAEP Redesign Project: Assessment Design Is the Key to Useful and Stable Assessment Results. Working Paper Series.

INSTITUTION American Coll. Testing Program, Iowa City, IA.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
REPORT NO NCES-WP-97-30
PUB DATE 1997-10-00
NOTE 120p.
AVAILABLE FROM U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Avenue, N.W., Room 400, Washington, DC 20208-5654.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Data Analysis; *Data Collection; Educational Assessment; Elementary Secondary Education; Equated Scores; *Estimation (Mathematics); National Surveys; *Research Design; Research Methodology; Sample Size; *Sampling; Scaling; Standards; Trend Analysis
IDENTIFIERS American College Testing Program; *National Assessment of Educational Progress; *Weighting (Statistical)

ABSTRACT

This report presents an investigation by the American College Testing Program (ACT) of an alternative design for the National Assessment of Educational Progress (NAEP). The proposed design greatly simplifies the data collection and analysis procedures needed to produce assessment results and has the potential to produce results that are more timely and easier to interpret. The plan calls for developing individual NAEP forms, where each individual form represents, as closely as possible, the assessment questions from the domain of knowledge being measured by an NAEP construct. Sets of these forms could be administered, in random order, to students in the schools. This would replace the balanced incomplete block design (BIB) currently used. Assessments constructed under the BIB design do not closely represent, at least for the 1996 science assessment, the content framework. Enhanced procedures are also suggested for developing precise content and statistical specifications for individual forms and procedures for pretesting items. The basic scores that ACT suggests using for producing group assessment results are calculated by weighting item scores from multiple-choice and constructed-response items, where the weights are determined, a priori, by content specialists. These weights should relate more closely to the weighting intended by content specialists than do current NAEP weights. Scaling, equating, and score distribution estimation methods are described that rely on less stringent psychometric and statistical assumptions than do current procedures. Issues in sampling that include sample size requirements, sample design, and estimating standard errors are also examined, as are procedures for reporting score distributions that reflect group performance on content domains. The feasibility of using such domain scores to measure trends and to facilitate setting NAEP standards is

+++++ ED417199 Has Multi-page SFR---Level=1 +++++
explored. ACT recommends focusing on the design of assessments and the data collection methods rather than on complex analysis procedures. (Contains 18 tables, 9 figures, and 41 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 417 199

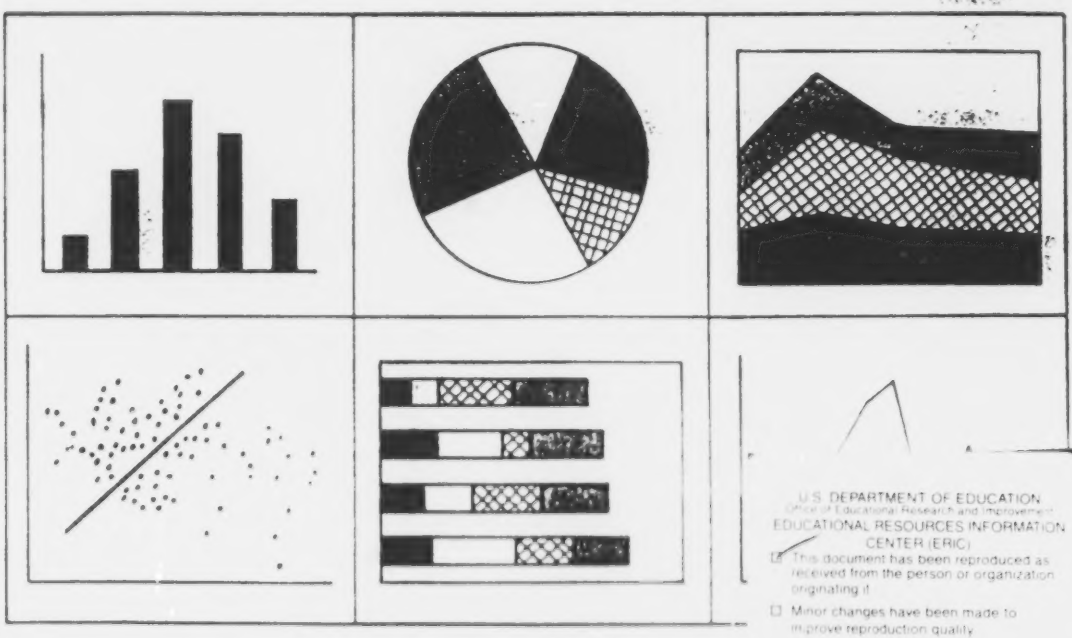
NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

ACT's NAEP REDESIGN PROJECT:

Assessment Design is the Key to Useful and Stable Assessment Results

Working Paper No. 97-30 October 1997



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

U.S. Department of Education
Office of Educational Research and Improvement

TM028140

BEST COPY AVAILABLE

ACT's NAEP REDESIGN PROJECT:

Assessment Design is the Key to Useful and Stable Assessment Results

Working Paper No. 97-30 October 1997

Contact: Steven Gorman
Assessment Group
(202) 219-1937
e-mail: steven_gorman@ed.gov

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Ricky T. Takai
Acting Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

Assessment Group
Gary W. Phillips
Associate Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, DC 20208

Suggested Citation

U.S. Department of Education. National Center for Education Statistics. *ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results*, Working Paper No. 97-30, by Luz Bay, Lee Chen, Bradley A. Hanson, Jay Happel, Michael J. Kolen, Timothy Miller, Mary Pommerich, James Sconing, Tianyou Wang, and Catherine Welch. Project Officer, Steven Gorman. Washington, D.C.: 1997.

October 1997

Foreword

Each year a large number of written documents are generated by NCES staff and individuals commissioned by NCES which provide preliminary analyses of survey results and address technical, methodological, and evaluation issues. Even though they are not formally published, these documents reflect a tremendous amount of unique expertise, knowledge, and experience.

The ***Working Paper Series*** was created in order to preserve the information contained in these documents and to promote the sharing of valuable work experience and knowledge. However, these documents were prepared under different formats and did not undergo vigorous NCES publication review and editing prior to their inclusion in the series. Consequently, we encourage users of the series to consult the individual authors for citations.

To receive information about submitting manuscripts or obtaining copies of the series, please contact Ruth R. Harris at (202) 219-1831 or U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Ave., N.W., Room 400, Washington, D.C. 20208-5654.

Samuel S. Peng
Acting Director
Statistical Standards and Services Group

ACT'S NAEP REDESIGN PROJECT:

Assessment Design is the Key to Useful and Stable Assessment Results

Prepared by:

Luz Bay
Lee Chen
Bradley A. Hanson
Jay Happel
Michael J. Kolen (Project Leader)
Timothy Miller
Mary Pommerich
James Sconing
Tianyou Wang
Catherine Welch

Prepared for:

U.S. Department of Education
Office of Educational Research and Development
National Center for Education Statistics

October 1997

Acknowledgments

We wish to thank the people on our Advisory Committee: Robert Brennan, Leon Burmeister, Mark Reckase, and Richard Sawyer. Christina Aicher and Sarah Logan both provided invaluable secretarial support.

Table of Contents

Foreword		iii
Acknowledgments		vi
Abstract		ix
Chapter One	INTRODUCTION	1
	Overview of ACT's NAEP Redesign Approach	
	Relationship of ACT's Approach to the Overall NAEP Redesign Effort	
	Preview	
Chapter Two	TEST DEVELOPMENT	9
	Content Specifications	
	Assumptions	
	Process	
	Discussion	
Chapter Three	PSYCHOMETRIC DESIGN	17
	Scoring, Scaling, and Equating	
	Estimating Distributions	
	Advantages and Disadvantages of the Proposed Design	
	An Investigation of Context Effects Based on Item-Block Position on the 1996 NAEP Science for Grade 8	
	Summary	
Chapter Four	SAMPLING METHODOLOGY	57
	Sampling Methodology	
	Proposed Changes to Sampling Methodology	
	Rationale for Proposed Changes	
	Simulation to Compare Methodology	
	Results	
	State and National Assessments	
	Discussion	
Chapter Five	FURTHER PSYCHOMETRIC ANALYSES	65
	Projecting to a Domain	
	Setting Standards for NAEP Via Domain Scores	
	Measuring Trends Over Time Via Domain Scores	
	Incorporating Background Variables Into Analyses	
	Conclusions	
Chapter Six	SUMMARY AND CONCLUSIONS	97
	Assessment Design	
	Scoring	
	Scaling and Equating	
	Estimating Score Distributions	
	Sampling	
	Further Psychometric Analyses	
	Independent Replication	
	Conclusions	
References		101

ABSTRACT

This report presents an investigation by ACT of an alternative design for the National Assessment of Educational Progress (NAEP). The proposed design greatly simplifies the data collection and analysis procedures needed to produce assessment results. The design has the potential to provide more timely results by using procedures that are less complex and that are easier to understand than are current procedures. The design also produces results that can be readily interpreted and replicated by other psychometricians and assessment agencies. The recommended procedures are intended to make it possible to describe NAEP more clearly to the assessment and educational communities, to lawmakers, and to the general public.

A plan is presented for developing individual NAEP forms, where each individual form represents, as closely as possible, the assessment questions from the domain of knowledge being measured by a NAEP construct. If necessary to fully represent the content framework, sets of these individual forms could be administered, in random order, to students within schools. This assessment design would replace the balanced incomplete block (BIB) design that is currently used. As is shown in this report, the assessments constructed under the current BIB design do not, at least for the 1996 NAEP Science Assessment, closely represent the content framework. In fact, we were able to develop only one individual form from the entire set of test questions that comprise the 1996 NAEP Science booklets. In addition, the BIB design requires strong statistical assumptions that likely do not hold well in practice. We suggest enhanced procedures for developing precise content and statistical specifications for individual forms and procedures for pretesting items. Although our recommended design constrains the test development process to a greater extent than it is now, our design can lead to assessments that overcome many of the difficulties and complexities inherent in the current design.

The basic scores that ACT suggests using for producing group assessment results are calculated by weighting item scores from multiple-choice and constructed-response items, where the weights are determined, *a priori*, by content specialists. These weights should relate more closely to the weighting

intended by content specialists than do the current NAEP weights. Scaling, equating, and score distribution estimation methods are described that rely on less stringent psychometric and statistical assumptions than do current procedures. These methods have the potential to produce results that are more stable and easier to interpret than those from present methods. The procedures we recommend do not need to incorporate plausible values, complex estimation procedures, or background variables into the estimation of the reported score distributions.

ACT firmly believes that all important NAEP results should be independently replicated. Currently, NAEP scaling and distribution estimation procedures incorporate heuristics that are not described in sufficient detail for independent replication. ACT recommends that this situation be remedied by using much simpler psychometric and statistical methods that are improvements over current procedures, and by making public all algorithms that are used in the scaling and distribution estimation process.

In ACT's design, issues in sampling that include sample size requirements, sample design, and estimating standard errors are also examined. We also study procedures for reporting score distributions that reflect group performance on content domains. The feasibility of using such domain scores to measure trends and to facilitate setting NAEP standards is explored. We investigate the use of multilevel models as a means for incorporating background variables in domain score estimation, and whether alternate sampling designs could be implemented for score reporting at the state level.

ACT strongly recommends focusing on the design of the assessments and the data collection methods, rather than on complex analysis procedures. The core concept of this design, which derives from ACT's many years of developing educational assessments, is that assessment design is the key to useful and stable assessment results.

ACT welcomes feedback from the National Center on Education Statistics (NCES), and would be pleased to work closely with NCES to refine and expand upon these ideas.

CHAPTER ONE
INTRODUCTION

This project provides ideas for redesigning the National Assessment of Educational Progress (NAEP) that simplify statistical and psychometric analyses and make the results of NAEP more useful to policymakers, educators, and the public. The objectives of the project are the following:

(a) To devise a process for test development that leads to tight test content specifications, statistical specifications, and tests;

(b) To reduce the reliance on strong psychometric assumptions in scaling;

(c) To simplify scaling and equating analyses;

(d) To simplify the sampling procedures used; and

(e) To develop procedures for enhancing score reporting.

ACT welcomes feedback from the National Center for Education Statistics (NCES). ACT would be pleased to work closely with NCES to refine and expand upon these ideas, and to focus on those ideas that are most likely to benefit the NAEP redesign effort.

Overview of ACT's NAEP Redesign Approach

ACT has designed procedures that rely, much more than does the current NAEP, on rigorous test development procedures, including well-defined content and statistical specifications and on more straightforward data collection designs for scaling and sampling. ACT believes that NAEP results can be produced using less complicated analysis procedures and with less stringent statistical and psychometric assumptions than those currently used. These less complex procedures could be more readily interpreted and replicated by other psychometricians and testing agencies. The use of less complex procedures should also make it possible to describe NAEP more clearly to the measurement, assessment, and educational communities.

Currently, there is delay in reporting scores from NAEP. Part of the reason for this delay is the extensive statistical and scaling analyses that are required before score data are reported. ACT believes that through careful test design, design of data collection and sampling procedures, and reliance on straightforward scaling analyses, some very useful results could be reported to NCES within one month following test scoring. These less complex procedures could also reduce the time needed to review procedures, and could help avoid the time-consuming delays that can result when problems are encountered with more complex procedures. Results that require more analysis could then be reported later.

The procedures that we envision are based on the development of individual NAEP forms, where each form can be considered to represent, as closely as possible, the domain of knowledge being measured by a NAEP construct. For content areas that cannot be adequately assessed by an individual form, sets of forms would be administered, in random order, to students within schools.

The construction of individual forms requires development of precise content and statistical specifications. The individual form structure that we propose differs considerably from the balanced incomplete block (BIB) assessment structure currently used by NAEP. In our demonstration in Chapter Two with the 1996 NAEP Science Assessment, we found that it was possible to develop only one individual form from all of the items used in the BIB blocks in the 1996 NAEP Science Assessment. In addition, we found that the set of items in all of the BIB blocks actually used do not reflect the assessment content as described in the Science Framework. Considering these shortcomings of present NAEP test development procedures, the proposed development procedures can be expected to result in assessments that much more closely represent the assessment content as described by the Frameworks.

In our experience, the development of individual forms will be most successful if the items are pretested on a reasonably representative sample, prior to construction of final forms. Therefore, we recommend enhanced procedures for pretesting that facilitate construction of individual forms that meet the more stringent criteria that we believe are required for NAEP. Although the test development

procedures that we recommend constrain the test development process to a greater extent than it is now. we believe that our design can overcome many of the shortcomings inherent in the current design.

We also recommend using standard test administration and sampling procedures. Specifically, we suggest that the individual forms be randomly assigned within schools. If individual forms are developed and administered in this way, then it should be possible to use standard equating and scaling procedures. Such procedures can likely be accomplished without the need to incorporate plausible values, complex estimation procedures, or background variables into the estimation of the score distributions that are reported.

One of ACT's concerns with current NAEP administration procedures is that BIB spiraling requires item blocks to be administered in different physical positions in the test booklets. To the extent that the position of the item block affects item difficulty, error is introduced into the scaling and linking. The use of individual forms will result in constant item position, thus eliminating one potential source of systematic error in scaling. Eliminating this source of error would be expected to lead to more stable modeling and more precise measurement

In addition, in the current BIB spiraling, the booklet an examinee is administered consists of a set of blocks. The content and statistical characteristics of the booklets are not controlled as tightly as they would be for the individual forms we are recommending. To the extent that psychometric model assumptions, such as unidimensionality, are violated, systematic error could be expected to affect the scaling results and, ultimately, the norms that are produced. Because the proposed individual forms are precisely defined in content and statistical characteristics, the psychometric analyses would be expected to be less affected by violations of assumptions than would analyses conducted under the BIB spiraling design.

We investigate psychometric model alternatives to the item-level psychometric modeling currently used with NAEP. Modeling the assessment data at the form level as we recommend, rather than at the item level, also can lead to more robust psychometric analyses.

The relative weighting of multiple-choice and constructed-response item scores in the current NAEP design is based largely on psychometric considerations that result from application of item response models. One potential concern with the current procedures is that the relative weights given to multiple-choice and constructed response items might differ from the weights intended by policymakers and item format matter specialists. We believe that specifications for weighting should be embedded in the development of frameworks and specifications. We report on an investigation of procedures that consider both psychometric and content-matter information in deriving the weights. The main conclusion of this investigation is that because the psychometric properties are similar for different weightings, the weighting used that should be the most readily interpretable.

The psychometric procedures used for scaling the current NAEP are extremely complex. These scaling procedures incorporate complex heuristics that are not described in sufficient detail in technical documentation for independent psychometricians to replicate NAEP scaling results. In addition, much of the scaling software, and the associated computer code, used in the NAEP program is proprietary. These complexities make it impossible for psychometricians elsewhere to fully judge the adequacy of the scaling that is done or the norms that are estimated. This combination of complexity and the inability of external researchers to replicate results creates an extremely unhealthy situation for NAEP. Even if all of the analyses are done flawlessly, lack of independent replicability could lead some individuals to question the accuracy and usefulness of NAEP results. To help avoid these potential problems, ACT proposes to use much simpler analyses that can be replicated by external researchers. ACT strongly recommends that all important NAEP results be independently replicated.

In this report, we also explore alternatives to the current extensive reliance on background variables in the analyses. One process often used in sampling is to oversample specific target groups (e.g., racial/ethnic groups). We explore the extent to which oversampling can be used more than it is presently, to reduce the reliance on background variables. We also recommend using school as the primary sampling unit, that schools be selected with equal probability, and that all students within school be tested. We

describe a plan for combining samples from the State and National Assessments. We recommend procedures that will likely facilitate reporting distributions of scores that represent performance on a domain, reporting score distributions at the state level, estimating trends, and incorporating background variables into the domain score analyses using multilevel models. We also recommend how ACT could further develop these ideas based on our general approach.

We use the current NAEP Science Assessment to demonstrate ACT's approach. We chose the NAEP Science Assessment for our demonstration because it contains a mix of multiple choice and constructed response items, with a large portion being polytomous. Also, ACT has many years of experience in constructing science tests and considerable expertise in this area. To demonstrate our approach, we use item parameter estimates and data from the most recent NAEP Science Assessment.

We wish to make it clear that we do not intend to recommend that NAEP scores be reported at the individual level. Even with the procedures we describe, the individual forms might not support interpretation of scores from individuals.

Relationship of ACT's Approach to the Overall NAEP Redesign Effort

ACT's NAEP Redesign project addresses the following Invitational Priority listed in the December 13, 1996 Federal Register:

> *Invitational Priority 4 - Psychometric procedures that maximize test reliability while minimizing analytic complexity and processing time.*

This project also addresses the following two areas presented in the Letter to Prospective Applicants dated December 20, 1996:

6. *Use of innovative psychometric procedures to calibrate, scale, score, link, and analyze NAEP data.*

7. *Development of analysis and reporting techniques that provide the public an initial release of results on a timely and predictable schedule.*

ACT's approach is consistent with and follows from many of the ideas described in the following three reports:

1. *Design/Feasibility Team Report to the National Assessment Governing Board (NAGB)* (Forsyth, Hambleton, Linn, Mislevy, & Yen, 1996);

2. *Operational Vision for NAEP - Year 2000 and Beyond* (NCES, 1996);

3. The *NAGB Policy Statement on Redesigning the National Assessment of Educational Progress* (NAGB, 1996).

To develop this approach, we have combined many of the ideas described in these reports with our experience in test development and analysis. The following statement from NAGB (1996) is one factor that led us to our approach.

> *The current National Assessment design is overburdened, inefficient, and redundant. It is unable to provide the frequent, timely reports on student achievement the American public needs. The challenge is to supply more information, more quickly, with the funding available. (p. 2)*

Those reports all discuss the concept of marketbaskets. Our ideas are closely related to this marketbasket concept. Specifically, the concept of an *individual form* that we define later in this report is very similar to the concept described by Forsyth et. al (1996): *Variation #1: Marketbasket is the Size of a Typical Assessment Form* (pp. 6-28, 6-29). The concept of a *super form* is very similar to the concept of *Variation 2: Marketbasket Larger than a Typical Assessment Form*. The concept we refer to as a *content domain* and scores that reflect that domain are closely related to the concept described in Forsyth et al. (1996) *Variation #3: Marketbasket Constitutes Subject Domain*. We chose to use our own terminology in the present approach because of our focus on precise content and statistical specifications for individual forms and domains. We believe that the general concepts we are discussing are consistent with the premises described in NAEP (1996). The concepts are also consistent with the objectives stated in NAGB (1996).

ACT bases its approach on its experience in developing educational achievement tests over the past 38 years. ACT has considerable experience in scaling and equating its own tests, as well as various

certification, licensure, and other tests on contract. One of the hallmarks of ACT's procedures is the ability to quickly produce results that are of high quality. For example, each year ACT administers and equates approximately 8 new forms of the ACT Assessment. The equating process is done without any delay in score reporting; score reports are mailed to examinees within 3 weeks of test administration. ACT believes that many of the test development and statistical analysis processes that we use can be applied to NAEP to allow results to be reported more quickly.

In this project, we present ideas for NAEP redesign in some of the areas that we believe improvements can be made. ACT is committed to contributing to the NAEP redesign effort, and we offer to make available our considerable experience and expertise in educational measurement, sampling, statistical analysis, and reporting. We will work cooperatively with NCES to ensure that our work contributes significantly to the redesign of NAEP.

Preview

The remaining chapters of this draft report present ACT's approach in more detail. Chapter Two presents ACT's redesigned test development procedures. We demonstrate these procedures using the 1996 NAEP Science Assessment. Chapter Three presents ACT's redesigned psychometric procedures. We focus on issues in weighting, scaling, and estimating score distributions. We also provide a conceptual framework for comparing ACT's redesigned development and psychometric procedures to present NAEP procedures. Chapter Four presents ideas for redesigning sampling procedures. In Chapter Five, we describe and evaluate a procedure for estimating domain scores for group-level data. We describe how these domain scores can be used to assess trends over time, and discuss how background variables might be used to make domain score estimation more precise. Chapter Six provides a summary of ACT's redesign approach along with conclusions and recommendations.

CHAPTER TWO
TEST DEVELOPMENT

This chapter evaluates the extent to which parallel forms could be constructed from existing NAEP item pools. The use of parallel forms that adequately cover the content and that could be completely administered to individual examinees in the amount of assessment time deemed feasible would lead to considerable simplifications in the administration design and analyses as suggested in Chapter One and as discussed fully in Chapter Three. The work presented here is provided as an illustrative example and employs the current NAEP eighth grade science assessment pool. The first step in this process consisted of consolidating the various types of specifications (content, cognitive, themes, and item formats) presented in the framework into specifications for an individual form. Then, the pool was evaluated for its capability to provide sufficient numbers of items to meet the various components of the specifications. Items were then selected and an initial form was constructed. This form represents a single assessment that can be used to evaluate the NAEP science domain while remaining consistent with framework recommendations concerning item formats and time restrictions. This form is not intended to represent one of the forms used in the current NAEP administration design.

The NAEP rationale and framework for the 1996 NAEP Science Assessment was presented in considerable detail in *NAGB (1996)*. As is described in the present chapter, the NAEP Science Framework presents a detailed description of the content structure of the NAEP Assessment. In this chapter, we provide an overview of the framework, and show how the 1996 NAEP science item pool relates to the *Science Framework for the 1996 NAEP*. (As it turns out, the content distribution of items in the pool is inconsistent with the framework in important ways). In our demonstration, we have assumed that the *Science Framework* provides the best representation of the content, as described by the subject matter experts who designed the framework. For this reason, we have attempted to build a single NAEP form that best represents the framework. Careful attention was paid to both the content and technical specifications in the construction of this single form. ACT recognized the importance of the content

specifications set forth in the framework and recommended that the content specifications be the first priority of this project. However, in addition to content specifications, an iterative development process was followed that attempted to construct a technically sound assessment on which NAEP achievement levels could be set.

Content Specifications

The *Science Framework for the 1996 NAEP* identifies four dimensions over which students shall be assessed. Several domains are identified within each dimension, and the framework provides guidelines as to how much of the item pool at each grade shall address each domain. The dimensions, domains, and guidelines for the construction of the item pool are presented in Table 2.1.

Table 2.1 Guidelines for Construction of the NAEP Science Assessment

Dimension	Domains	Percent of Testing Time
Content	Physical Science (PS)	30
	Earth Science (ES)	30
	Life Science (LS)	40
		100%
Cognitive	Scientific Investigation (SI)	30
	Practical Reasoning (PR)	25
	Conceptual Understanding (CU)	45
		100%
Nature of Science	Nature of Science (NS)	**
	Nature of Technology (NT)	**
		15%
Themes	Patterns of Change (PC)	**
	Models (MOD)	**
	Systems (SYS)	**
		50%

** The 1996 Science Framework is organized according to content and cognitive areas. In addition, the Nature of Science and Themes are categories that integrate the three content areas. For example, 15 percent of the content should measure the Nature of Science and 50 percent should assess understanding of the themes, spread evenly across all three themes.

BEST COPY AVAILABLE

In addition to content specifications, the framework contains guidelines concerning the total amount of testing time for each item type. For example, multiple-choice items should comprise no more than 50 percent of the assessment, as measured by student response time. In addition, open-ended items should make up at least 50 percent of the assessment, as measured by student response time. About one-third of the open-ended questions should consist of extended response items (NAGB, 1996, pg 43). There are no guidelines with regard to crossing content and cognitive specifications with nature of science or themes categories.

Assumptions

Several assumptions concerning testing time were taken as given prior to test construction. First, it was assumed that the total testing time would be 90 minutes. Second, within the 90-minute framework, given the guidelines presented in the framework and the *Science Assessment and Exercise Specifications for NAEP* document, ACT designated 30 minutes (approximately 30%) to be reserved for a hands-on exercise, 45 minutes (approximately 50%) multiple choice and dichotomously-scored open-ended items, and 15 minutes (approximately 20%) short-constructed and extended-constructed items. Finally, testing times for each item type were assumed as follow: multiple choice and dichotomously-scored open-ended items (1 minute), short-constructed items (2 minutes) and extended response (5 minutes).

Process

The initial step in forms construction involved assessing the current pool with respect to the content and cognitive classifications. Attention focused on the marginal counts of items in the two-way, content-by-cognitive classification table and on stratification of this table by item type. A form was then constructed to meet the specifications while simultaneously attempting to meet guidelines with respect to testing time by item type. Finally, in constructing the forms, it was recognized that certain blocks needed to be held together because of common stimulus materials. This presented a problem because it

eliminated several items from consideration. While such blocks could be used in a matrix design, for the current approach they would, by their size, lead to important deviations from the specifications.

The current pool for eighth grade science contains 190 items with classification data on at least one of the dimensions. Table 2.2 contains the two-way classifications.

Table 2.2 Number of Items in Each Two-Way Classification Cell for the 1996 NAEP Science Pool

	Scientific Investigation	Practical Reasoning	Conceptual Understanding	Not Classified	Total
Physical Science	13	17	31	1	62
Earth Science	12	12	34	5	63
Life Science	9	14	42	0	65
Total	34	43	107	6	190

* Using assumed testing times

Examination of the pool with respect to the constraints given in Table 2.1 and the actual number of different item types given in Table 2.2 made it clear that only one complete form could be constructed that would meet all constraints. Part of the problem was the limitations created by the need to hold together certain blocks. After those blocks were used, many of the content and cognitive categories were virtually depleted after the construction of a single form.

The results for a single form that best represents the framework are given in Table 2.3. The results for this form indicated close correspondence to the frameworks. The match-to-content specifications was very close. The match to the cognitive specification was a somewhat weaker, owing to some extent to an over-representation of conceptual understanding items in the pool and an undersupply of scientific investigation items. It was also difficult to match the requirement of 50 percent themes items evenly distributed among the three domains (patterns of change, models and systems); again, this reflected some disproportionate percentages in the pool.

Table 2.3 1996 NAEP Grade 8 Science ACT-Constructed Forms*

	M	Dich. OE	SCR	ECR	Sum	% of All Items	Time (min.)*	% of Total Time	*Specs.*
Content (includes items from HO block)									
Physical Science (PS)	12	2	2	1	17	29.8%	23	31.1%	*30%*
Earth Science (ES)	13	2	4	0	19	33.3%	23	31.1%	*30%*
Life Science (LS)	14	3	3	1	21	36.8%	28	37.8%	*40%*
Total	39	7	9	2	57	100%	74**	100%	
Cognitive (includes items from HO block)									
Scientific Investigation (SI)	4	0	2	2	8	14.0%	18	24.3%	*30%*
Practical Investigation (PR)	7	3	4	0	14	24.6%	18	24.3%	*25%*
Conceptual Understanding (CU)	28	4	3	0	35	61.4%	38	51.4%	*45%*
Total	39	7	9	2	57	100%	74**	100%	
Themes (includes items from HO block)									
Patterns of Change (PC)	2	2	4	1	9	15.8%	17	23.0%	*16.7%*
Models (MOD)	7	1	0	0	8	14.0%	8	10.8%	*16.7%*
Systems (SYS)	6	2	4	0	12	21.1%	16	21.6%	*16.7%*
Total	15	5	8	1	29	50.9%	41	55.4%	*50%*
Nature (includes items from HO block)									
Nature of Science (NS)	2	1	2	1	6	10.5%	12	16.2%	*9%*
Nature of Technology (NT)	4	1	1	0	6	10.5%	7	9.5%	*6%*
Total	6	2	3	1	12	21.0%	19	25.7%	*15%*

* Using assumed testing times

** This total does not include HO time spent working with kit.

Testing Time*		
Multiple Choice (MC)	39	39 min. (excluding HO)
Dichotomous Open-Ended (Dich-OE)	6	6 min. (excluding HO)
Short Constructed Response (SCR)	5	10 min. (excluding HO)
Extended Constructed Response (ECR)	1	5 min. (excluding HO)

Item Format	**Time (min).***	**Total Time**	**Specs.**
MC/OE	45	50.0%	50%
SCR/ECR	15	16.7%	20%
HO	30	333%	30%
	90 > = *50 min.*		

The characteristics of the pool in terms of distribution percentages are summarized in Table 2.4. Some of the disproportionalities in evidence here are reflected in the constructed form.

Table 2.4 Characteristics of NAEP Science Pool*

Summary	Across All Items		Across All Booklets		
	% of Items	% of Time	% of Items	% of Time	*Specs.*
Content					
Physical Science (PS)	32.3%	34.4%	32.0%	34.1%	*30%*
Earth Science (ES)	33.3%	32.3%	33.8%	33.3%	*30%*
Life Science (LS)	34.4%	33.2%	34.2%	32.6%	*40%*
Cognitive					
Scientific Investigation (SI)	18.5%	23.1%	20.3%	25.0%	*30%*
Practical Reasoning (PR)	22.8%	26.9%	21.8%	25.2%	*25%*
Conceptual Understanding (CU)	8.7%	50.0%	57.9%	49.8%	*45%*
Theme					
Practical Reasoning (PR)	14.8%	15.6%	15.3%	15.8%	
Models (MOD)	12.7%	12.9%	13.1%	13.2%	
Systems (SYS)	21.2%	22.5%	21.2%	22.4%	
Total	48.7%	50.0%	49.6%	51.4%	*50% (evenly split)*
Nature					
Nature of Science (NS)	7.9%	12.3%	7.4%	10.9%	*9%*
Nature of Technology (NT)	7.9%	10.2%	7.3%	9.4%	*6%*
Total	15.8%	22.5%	14.7%	20.3%	*15%*

* Using assumed testing times

Discussion

These findings suggest that a complete form can be constructed that provides adequate content coverage in the allotted test time. However, it is clear that considerable item development would be needed to construct multiple parallel forms. Because the pool was virtually depleted to construct a single form, it would appear necessary to increase the pool to make it possible to construct parallel forms. Also, the use of nearly the entire pool to construct this form (excluding the blocks that needed to be held

together) meant that relatively little attention could be paid to the statistical characteristics of the form. Initially, in the process of selecting items, those with extreme difficulties or very low discriminations were passed over. However, it became apparent that if content specifications were to be met, statistical characteristics would need to take on a minor role in the construction process. This also has implications for item development. Ideally, the cells of the content-by-cognitive classification table should contain enough items to permit the development at that level of forms that are equivalent in both content and statistical characteristics. To the extent that multiple alternate forms depart from equivalence, the effectiveness of the equating procedures described in Chapter Three would be reduced.

In addition to increasing the scope of the item development, the forms construction process would benefit from the generation of stable item statistics that are good predictors of final-form performance. One model would allow item formats and various item types to be initially tested during a preliminary pilot administration. This pilot administration would identify the types of items and item formats that are most successful for the population of interest. Following the pilot administration, items should be refined and additional items produced in preparation for a field test of the items. The field test administration would produce item-level statistics that would be used to construct the final forms. Items that are either too difficult or too easy and those that fail to discriminate between students would be eliminated from the available pool for construction.

In general, the statistical characteristics of the forms would need to be specified so that sufficiently precise measurement would be expected to occur across a range of ability that encompasses the locations of the standards that would be set. If IRT parameters are to be used, test information functions could guide the forms development without explicit constraints on the difficulty and discrimination levels of the items. Deficiencies in the pool that lead to the inability to match the information functions would need to be remedied. If classical statistics are employed, some explicit constraints should be made. Without some preliminary indication of the location of the standards, it is difficult to specify exactly what these may need to be. However, experience suggests that difficulty values between .2 and .8 and discrimination

indices greater than .4 will create tests that discriminate across a broad range of ability. More restrictive specifications involving statistical constraints nested within content classifications could be used to more closely prescribe the expected performance of the forms and improve their degree of parallelism. Advances in test generation software make possible the matching of complex sets of test specifications across multiple forms and may be appropriate for use in this context.

This chapter has presented the results of an ACT-constructed form of the NAEP science assessment using the current item pool. Two types of test specifications need to be used throughout any assessment construction of this type: content specifications and statistical specifications. The content specifications, based on the guidelines presented in the framework, need to specify both the topics to be covered by the assessment and the proportion of items to measure each topic. The content coverage and the number of items included in each form need to reflect the emphasis in the framework. If a single-form approach to NAEP was adopted, the basic structure of the forms would remain the same from administration to administration, thus making the scores comparable after equating. The statistical specifications for the form need to indicate the level of difficulty and the minimum acceptable level of discrimination of the items to be used. These characteristics need to be selected so that the tests will effectively differentiate between students in the achievement levels-setting process. Forms that match the statistical specifications will be expected to provide a high degree of measurement precision for the achievement levels.

CHAPTER THREE
PSYCHOMETRIC DESIGN

The core of the proposed design is the creation of specific content and statistical specifications from which individual forms can be built. These specifications, referred to as *form specifications*, will represent the content of the Frameworks but be much more precise. The form specifications will include a precise specification of the items that will appear on a form in terms of content, item format, and statistical characteristics.

For some examinations, such as the 1996 NAEP Science Assessment considered in Chapter Two, it may be possible to write form specifications in such a way that forms can be produced that are administered in full to each examinee. In cases in which this is not possible, more than one administered form will need to be developed so that the set of items on all administered forms meets the form specifications. The items on any single administered form would not completely meet the form specifications. The set of unique items across all administered forms will be denoted a *super form*. A super form is constructed to meet the form specifications. The administered forms would not be parallel forms with respect to the form specifications, but would be as close to parallel as possible.

The administered forms will be more nearly parallel if they contain as many common items as possible, with as few items that are unique to each form as possible. For example, if the specifications include a hands-on exercise in each of three content areas and it is possible to have only one hands-on exercise per form, then administered forms can be built that are identical except for having a different hands-on exercise.

The next section discusses scoring, scaling, and equating procedures for the proposed design. NAEP results are reported as properties of distributions of random variables. The second section discusses estimation of distributions used to report NAEP results. The third section discusses some advantages and disadvantages of the proposed design relative to the current design. An advantage of the proposed design is that much simpler analysis procedures, requiring fewer assumptions, can be used. When the assumptions

required by the complex analysis procedures used in the current design are violated invalid results can be produced (Zwick, 1991). One assumption required by the analysis procedures used in the current design is that the performance of an item is not influenced by the context in which the item is presented. The fourth section presents results based on data from the 1996 NAEP Science Assessment that provides some indication that the position of an item in a block influences performance on the item. The final section summarizes the crucial aspects of the proposed psychometric design.

Scoring, Scaling, and Equating

Scoring, scaling, and equating procedures initially will be described under the assumption that the specifications are written such that parallel forms, administered in full to examinees, can be built. A later sub-section will discuss scoring, scaling, and equating procedures when super forms are needed to adequately represent the test specifications.

As described in Chapter Two, a single form has been constructed for the eighth grade NAEP Science Assessment that appears to adequately cover the content in the Science Framework. In constructing this form, the NAEP Science Framework was used as the form specification (more precise specifications based on the Framework would be developed to actually implement the proposed design). This single form will be used for illustrative purposes in this chapter.

Scoring

For each NAEP examination, results are reported using a single scale. In the current NAEP design, this scale is defined by latent variable models for the item response data. For example, for the Science Assessment each item is assumed to measure one of three latent variables corresponding to three content areas. The scale used to report results is based on a linear combination of the three latent variables. In the proposed design, scores on a form or super form are the basic data used in the analyses to produce NAEP results. This necessitates determining an explicit procedure for translating scored item responses into a form score.

The influence of items on the scale in the present NAEP design is based largely on psychometric considerations that result from application of item response models. One potential concern with the current procedures is that the relative weights given to multiple-choice, dichotomous open-ended, short constructed response, and extended constructed response items might differ from the weights intended by content matter specialists. The proposed procedures consider statistical, psychometric, and content-matter information in deriving the weights.

In NAEP, as presently designed, the intended relative weighting of item types by content matter specialists is not fully developed in the frameworks or in the specifications documents. To help clarify this issue, consider the following simple example: Suppose we have a short test composed of three dichotomously scored multiple-choice items and one constructed response item that is scored on a 0 to 3 scale. Suppose also that constructed response and multiple-choice items are to be equally weighted. In terms of number of score points that contribute to the maximum summed score, the test equally weights the two types of items. In terms of testing time, the items likely are unequally weighted, because the multiple-choice items will likely take much less time than the constructed response item. Is the weighting by content specialists intended to be by numbers of score points, by testing time, or by numbers of items? Present NAEP framework and specifications documents do not provide a clear answer.

The numbers of score categories used by judges to score the constructed response items also can influence the weightings. The number of categories, and the numerical values assigned to each category (e.g., 0, 1, 2 versus 1, 2, 3) should also be part of the specifications. In present NAEP analyses, the score categories on constructed response items are often collapsed so that the algorithms used in NAEP scaling procedures will converge. Such collapsing of categories can affect the relative weighting given to the different item formats. In addition, the combining of score categories can make it difficult for content specialists and policy makers to understand the meaning of the scores, and can cause confusion when NAEP standards are set. The scaling and weighting procedures that ACT is recommending directly address these concerns. In addition, these procedures make explicit the relative weights of different item

types that are used so that NAEP standard setters can incorporate this information in the standard setting process.

It is proposed that a weighted combination of item scores on a form be used as the form score. These form scores would be an *a priori* weighted linear combination of the item scores. We suggest that the choice of item weights used to compute a form score be an integral part of constructing the form specifications. Content and format should be principal considerations in determining the weights.

Wang and Stanley (1970) indicate that, from a psychometric perspective, differential weighting is only effective when there are only a few measures in the composite that are not highly related to one another. This suggests that a small set of mutually exclusive unweighted sums of item scores first be formed based on content or item format. These sums will be denoted form subscores. Differential weighting to produce a form score would only be applied to the form subscores. Hence, weights do not need to be determined for individual items, but for a smaller number of sums of item scores.

There are three main factors to consider in determining weights: 1) item content, 2) item format, and 3) the statistical and psychometric properties of the resulting scores. Content is important to consider in weighting as it affects the meaning of the score. Presumably, the reason for having different item formats is that the different formats measure somewhat different things. The *a priori* weighting used to compute a form score will determine the influence of each of the constructs measured by the items in the score, and consequently the meaning of the score. Determination of weighting with regard to format and content should be made by policy-makers and/or content matter experts who are involved in designing the frameworks, the specifications, and in designing the forms. The *a priori* weighting used to compute a form score would determine the influence of each of the constructs measured by the items in the score, and consequently the meaning of the score.

Statistical and psychometric considerations in deciding on weights include the statistical influence of the components of a score on the score distribution and the effect of weights on the measurement properties of the score (including reliability and the conditional standard error of measurement of the

observed score given the true score). Content considerations in weighting are addressed based on the content and format of the items and do not require empirical data. Empirical data are needed to address statistical and psychometric considerations in weighting.

One statistical consideration in determining weights involves the extent to which components contribute empirically to the form score, as opposed to the contribution of the components implied by the nominal weights (nominal weights are the *a priori* weights chosen). Wang and Stanley (1970) define the effective weight of a component in a composite score as the contribution of that component to the variance of the composite. Consideration should be given to the effective weights of the components in the score as well as to the nominal weights based on format and content.

Consideration of reliability in determining weights is especially important when the components of the score have fairly different reliabilities. An example is the case of combining multiple-choice and constructed-response items (Wainer & Thissen, 1993).

We believe that specifications for weighting should be embedded in the development of frameworks and specifications. These specifications include determining which sets of item scores should be added together to produce subscores, and the weights to be used for combining subscores into a form score.

Data on the base form used to define the scale could be used to incorporate statistical and psychometric factors in the weighting, and also possibly to adjust the nominal weighting based on statistical and psychometric considerations. For example, if effective weights are quite different from the nominal weights, then some adjustment of the nominal weights may be required. Still, to the extent possible, we believe the nominal weights should drive the weighting process because they are more readily interpretable by content-matter specialists and policy-makers.

We propose the following sequence of steps in determining the weighting used to produce form scores.

1. Define a small number of form subscores that are unweighted sums of item scores. Subscores are based on item format. Each subscore contains a unique set of items of a particular item format (possibly a single item), and each item is included in one and only one subscore.

2. Decide on nominal weights for each subscore that will be used to produce the linear combination of subscores that will serve as the total form score.

3. Using data on the base form, examine the statistical and psychometric properties of the form scores defined by the weighting in step 2. The analyses would include examining the effective weights and the reliability of the form scores in relation to the reliability of the subscores.

4. Possibly adjust the nominal weights determined in step 2 based on the empirical results in step 3.

These steps assume the score categories for the items are predefined and are not collapsed. The subsets of items used for subscores should be defined so that sets of items could be developed for other forms such that the statistical and psychometric characteristics of subscores on alternate forms are approximately equivalent (after possibly equating the subscores), and the relationships among the subscores approximately equal across forms. The nominal weights would be chosen in step 2 based on the format of the items in each subscore.

The weights developed with the above four steps for the first (base) form would be used as nominal weights for all parallel forms built to the form specifications. This weighting assumes precise form specifications that include the number of items of different response types that are in particular content categories. For example, the specifications might call for one extended constructed-response item that has life sciences content. The item used for this specification would differ in different forms, but be weighted the same in computing the form scores for each form. The item would be included in a particular subscore which would be weighted to produce the form score.

As an example of the effects of weighting we will look at some alternative weightings to produce scores for the 1996 NAEP Science form that ACT developed in Chapter Two. Two subscores will be

defined. One subscore is the sum of the item scores on the multiple-choice items (objectively scored) and the other subscore is the sum of the item scores on the constructed-response items (subjectively scored). Of the 57 items on the form, 39 items are multiple-choice and 18 items are constructed-response. The sum of the item scores on the 39 multiple-choice items range from 0 to 39, and the sum of the item scores on the 18 constructed-response items range from 0 to 31. Four weightings of the two subscores are considered. The first weighting is just a sum of the two subscores (weighting by number of score points). The second weighting is by the time allotment for the items (each subscore is converted to a proportion of possible points and multiplied by the time allotted for all items in the subscore). There are 39 minutes allotted to the multiple-choice items and 35 minutes allotted to the constructed-response items. These are the sums of times allotted to the individual items of each type; the time actually spent on an item may be different from the allotted time. The third weighting is by the number of items (each subscore is converted to a proportion of possible points and multiplied by the number of items in the subscore). The fourth subscore is the sum of the proportion of possible score points on the multiple choice items and the proportion of possible score points on the constructed-response items (equal weighting of multiple-choice and constructed-response item scores).

Item response data were simulated using the item response model used in the current design. The item parameters used were those that were estimated in the operational scaling for the 1996 NAEP Science Assessment. The distribution of the three latent variables was taken to be multivariate normal with means all equal to zero and standard deviations all equal to one. The correlations among the three variables were those estimated in the operational 1996 NAEP scaling for Science. The three latent variables for each examinee were generated from a multivariate normal distribution. For each simulated examinee, item responses for each item in the form built by ACT were simulated. A sample of 2000 examinees was generated. For each simulated examinee, the multiple-choice and constructed-response form subscores were calculated. These two subscores were used to calculate the four possible form scores.

The correlation between the multiple-choice subscore and the constructed-response subscore in the simulated data was .76. The correlations between the four form scores were all greater than .99. This result is consistent with the conclusions of Wang and Stanley (1970) that in many cases weighting does not make much of a difference on the statistical properties of the score. However, weighting might have serious implications for how scores are interpreted. In addition, the weights that are used might influence the standards that are set in NAEP standard setting procedures, even if the scores are highly correlated as in this example.

Table 3.1 Nominal and Effective Weights for Multiple-Choice (MC) and Constructed-Response (CR) Subscores

	Nominal Weights		Effective Weights	
Form Score	MC	CR	MC	CR
Weight by points	.50	.50	.58	.42
Weight by time	.47	.53	.54	.46
Weight by number of items	.63	.37	.72	.28
Equal weight for proportion of score points	.44	.56	.51	.49

The nominal and effective weights of the multiple-choice and constructed-response subscores for the four form scores are given in Table 3.1. Both the nominal and effective weights have been standardized to sum to one. The effective weights represent the contribution of the multiple-choice and constructed-response subscores to the variance of each form score. The effective weights are greater than the nominal weights for the multiple-choice subscore, and the effective weights are less than the nominal weights for the constructed-response subscore.

To assess the measurement properties of the four form scores the squared correlation of the observed and true scores were computed for each weighting over the 2000 simulated observations. True scores were computed for each examinee using the estimated parameters in the model used to simulate the data. These squared correlations are estimates of the reliability of the scores produced by the various weightings. The squared correlation was 0.83 for the score based on weighting by the number of items

in each subscore, and the correlations were 0.85 for the other three weightings. Thus, the reliability of the scores was very similar across the four weightings. The squared correlation between the multiple choice subscore and its true score was .77. The square of the correlation between the constructed response subscore and its true score was .71. The purpose of computing these reliabilities was to examine the effect of the various weightings on the measurement properties of the scores. The reliability of individual scores is not directly relevant for NAEP since scores for individual examinees are not reported.

In the proposed design the issue of how item scores are used to produce the scale on which results are reported needs to be explicitly considered. In the current design the latent variable model used for scaling determines the how the responses of students to individual items influence the scale. Issues of scoring can sometimes enter into analyses in the current design. For instance, there are cases in the current design in which the scoring of an item needs to be changed because the item response model used to produce the scale does not work with the item as originally scored. We believe that an explicit consideration of the influence of items on the scale helps to clarify the meaning of the scale, and may make results easier to interpret.

Scaling and Equating

When parallel forms can be developed that meet the specifications and can be completely administered to examinees, it is proposed that standard scaling and equating procedures be used with the form scores. If subscores are used that are linearly combined to create a form score (as described in the Scoring sub-section), these subscores might be equated before equating the form score. Alternatively, it might be better to equate scores at the form level. A scale could be defined based on the form score for a base form. As indicated by Petersen, Kolen, and Hoover (1989), the score scale could be defined using normative data, score precision considerations, or by incorporating information from test content.

One possibility in a design such as this is to administer a single form to most examinees on a particular administration. The form administered to most examinees will be denoted the *major form*. The major form would be either a form that had been previously equated to the NAEP score scale or the form

used to construct the score scale. In addition, one or more other forms and the major form would be given to randomly equivalent groups of examinees. This process allows all forms to be put on the same scale. The major form could be completely released, and the other forms used in future administrations to put new forms on the same scale. It is recommended that a random groups equating design be used to equate other forms to the major form. One method of equating that could be used is equipercentile equating with smoothing (Kolen & Brennan, 1995).

Simple and straightforward scaling and equating procedures using form scores can be used in the proposed design. These straightforward procedures contrast with the approach used in the current NAEP design, in which scaling and equating is accomplished with much more complex analyses using item response models with item data.

Scoring, Scaling, and Equating for Super Forms

A super form consists of a collection of items on a group of administered forms, where each administered form is given in full to examinees. The items in a super form meet the specifications for the examination, whereas the items in each administered form do not fully meet the specifications. Super forms are needed when forms that meet the specifications and are completely administered to examinees cannot be built.

We have proposed that a single form score be the basis of analyses to produce NAEP results. Super form scores cannot be directly computed because no single examinee takes a complete super form, so more complex analysis procedures will be required. When super forms are needed we recommend that the majority of items in each administered form be common among all administered forms. The common items should be as representative as possible of the complete specifications. The combination of common items and unique items on each administered form would comprise the super form.

The strategy we recommend for obtaining results for super forms is to first estimate distributions of super form scores (this will require the application of some appropriate models and assumptions). The

analysis procedures outlined in the preceding sub-section, where a single administered form meets the specifications, could then be used with the estimated super form distributions.

We will illustrate some potential procedures for estimating super form score distributions using the 1996 NAEP Science Assessment. The ACT NAEP Science form described in Chapter Two has 57 items, 6 of which are part of a single hands-on exercise. This hands-on exercise has 4 physical science items and 2 life science items. We constructed a super form by adding another hands-on exercise to the original form. The hands-on exercise added consists of 8 earth science items (there were no earth science items in the hands-on exercise in the original form). There are two administered forms. One form is the original ACT constructed form (denoted administered form 1). In the second administered form the hands-on exercise in the original form is replaced by the earth sciences hands-on exercise (this form is denoted administered form 2). Administered form 1 has 57 items, administered form 2 has 59 items, and there are 51 items common to the two forms. The super form consists of 65 items (the 51 common items, plus the 6 items on the first hands-on exercise, plus the 8 items on the second hands-on exercise).

The score of interest on the super form is the sum of item scores on the 65 items. The super form score ranges from 0 to 84. The score on the common items range from 0 to 58, and the scores on the hands-on exercises in administered forms 1 and 2 range from 0 to 12 and 0 to 14, respectively. Simulated responses to the 65 item were generated for 6000 examinees using parameters of the item response models used in the current design for the 1996 Science Assessment. The distribution of the three latent variables used in the item response models was multivariate normal with the correlations as estimated for the 1996 Science Assessment. The data for the 6000 examinees on all 65 items are referred to as the complete data. The first 3000 examinees were designated to take administered form 1 (containing only the first hands-on exercise), and the second 3000 examinees were designated to take administered form 2 (containing only the second hands-on exercise). Thus, in the analysis to compute a distribution of super form scores for the 6000 examinees, the responses of the first 3000 examinees to the second hands-on exercise, and the responses of the second 3000 examinees to the first hands-on exercise were not used. Half of the

examinees were treated as missing the responses to items in the first hands-on exercise, and the other half of the examinees were treated as missing the responses to items in the second hands-on exercise. The data with each examinee missing responses to one of the hands-on exercises are called the incomplete data.

Let X be the score on the common items, Y_1 be the score for the first hands-on exercise, and Y_2 be the score for the second hands-on exercise. All three variables are discrete, so the full data is contained in a three-way table (common item score by hands-on score 1 by hands-on score 2). Each cell in the three-way table contains the count of the number of examinees who obtained a particular combination of the three scores.

Analysis procedures for missing data will be used to estimate the counts in the complete three-way table. The distribution of the super form scores can then be computed from the counts in the complete three-way table. Data are observed for the X-Y_1 two-way marginal table and for the X-Y_2 two-way marginal table. Some assumptions need to be made in order to produce estimates of cell counts in the complete X-Y_1-Y_2 three-way table.

The first step was to fit a polynomial loglinear models to each of the two observed two-way tables. The model for the X-Y_1 table has the form:

$$\log\,(m_{ij}) = \alpha_0 + \sum_{r=1}^{n_1} \alpha_{1r} i^r + \sum_{s=1}^{n_2} \alpha_{2s} j^s + \sum_{r=1}^{n_1} \sum_{s=1}^{n_2} I(r,s)\, \gamma_{rs} i^r j^s, \qquad (3.1)$$

where m_{ij} is the expected count for score i on X (i = 0,...,51) and score j on Y_1 (j = 0,...,12), and I(r,s) equals zero or one for each r and s indicating which of the γ_{rs} are allowed to be non-zero. Polynomial loglinear models such as that in Equation 3.1 have been successfully used in applications involving test score distributions (Rosenbaum & Thayer, 1987; Holland & Thayer, 1987; Hanson, 1991a; Livingston, 1993). A model analogous to that given in Equation 3.1 was used for the X-Y_2 two-way table. It was found that for the X-Y_1 two-way table a model with n_1 = 5, n_2 = 7, I(1,1) = 1 and I(r,s) = 0 for $r \neq 1$ and $s \neq 1$ fit the data well. For the X-Y_2 two-way table it was found that a model with n_1 = 5, n_2 = 6, I(1,1) = 1 and I(r,s) = 0 for $r \neq 1$ and $s \neq 1$ fit the data well.

The assumption made in order to estimate the counts in the three-way table using the incomplete data was that the following polynomial loglinear model holds:

$$\log(m_{ijk}) = \alpha_0 + \sum_{r=1}^{n_1} \alpha_{1r} i^r + \sum_{s=1}^{n_2} \alpha_{2s} j^s + \sum_{t=1}^{n_3} \alpha_{3t} k^t + \gamma_{110} ij + \gamma_{101} ik \,, \qquad (3.2)$$

where m_{ijk} is the expected count for score i on X, score j on Y_1, and score k on Y_2, and $n_1 = 5$, $n_2 = 7$, and $n_3 = 6$. This model incorporates the assumption that Y_1 and Y_2 are conditionally independent given X (Y_1 and Y_2 can still be associated when not conditioned on X). Maximum likelihood estimates of the model in Equation 3.2 can be found using the data in the X-Y_1 and X-Y_2 two-way tables using the EM algorithm (Dempster, Laird, and Rubin, 1977). Procedures described by Rindskopf (1992) could also be used to fit the model in Equation 2 using the incomplete data.

Using the incomplete data for the 6000 examinees, the model in Equation 3.2 was estimated using the EM algorithm. The estimated cell counts in the complete three-way table were used to compute an estimated super form distribution. Figure 3.1 contains this estimated super form distribution along with the actual super form distribution for the complete data on all 6000 examinees. Note that estimated distribution is Figure 3.1 is based only on data in the X-Y_1 and X-Y_2 two-way tables (this is the data that would be obtained for the two administered forms), whereas the observed distribution in Figure 3.1 is the actual super form distribution of scores for the 6000 examinees (using the complete data). For this example the model in Equation 3.2 worked well for estimating the super form score distribution from the incomplete data.

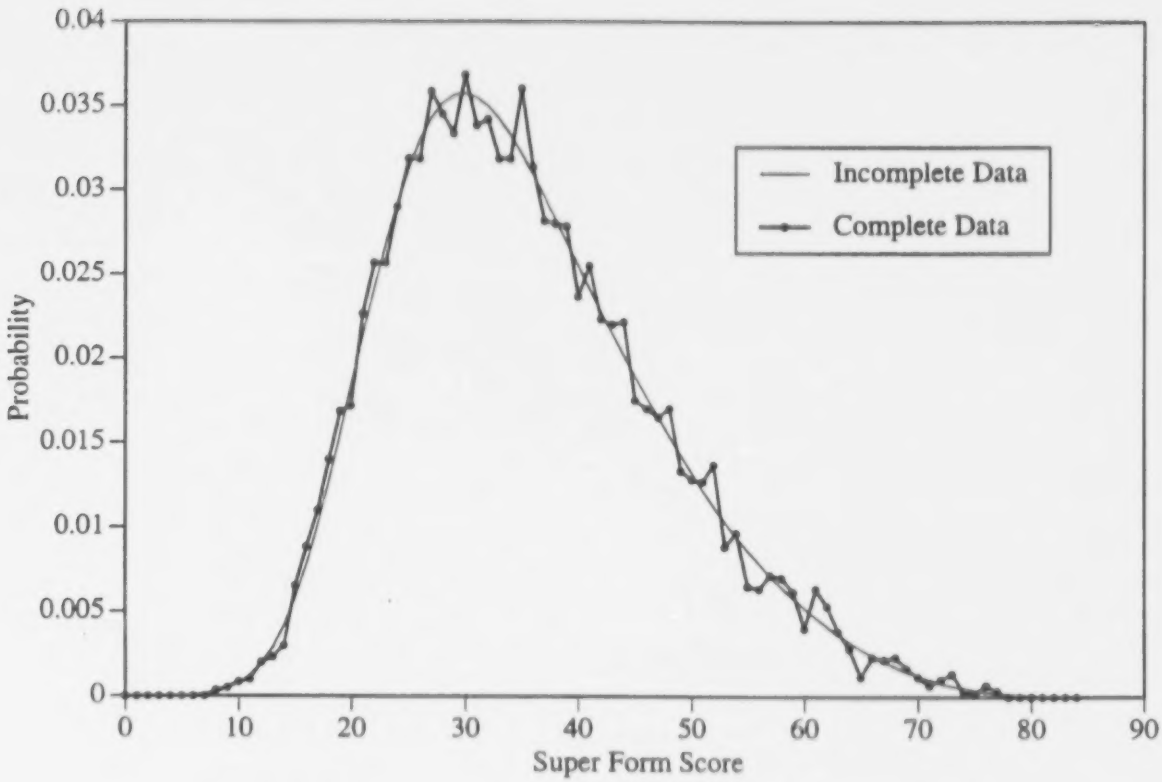


Figure 3.1 Distributions of super form scores observed for the complete data and estimated using the incomplete data

Other data collection designs are possible for the situation considered in the example. For instance, in addition to the two randomly equivalent groups of examinees who take administered forms 1 and 2, a third group taking only the two hands-on exercises could be used. This would provide some information on the association between the two hands-on exercises. The EM algorithm could be used to estimate the counts in the full table using this incomplete data and an appropriate loglinear model.

An alternative to the use of missing data methods as presented in the example is the use of latent variable models. Latent variable models derive their analytical power from the strong assumption that the observed variables are independent when conditioned on a small number of latent variables, and assumptions concerning the parametric form of functions giving the probability of item responses conditioned on the latent variables. The use of latent variable models may be most appropriate when results are to be reported using estimated latent variable distributions. Treating the analysis as a missing data problem is more straightforward when the goal is to estimate distributions of observed variables. The "observed" variables may not in fact be observed for any of the examinees, but they potentially could be observed. In the example, the super form scores were not observed for any examinee when using the incomplete data, but they potentially could be observed (as in the complete data).

An issue when specifying latent variable models for multiple observed scores is dimensionality, or the number of latent variables required for the observed variables to be conditionally independent given the latent variables. If a latent variable model were used in the example where there were three variables (a common item score, and two hands-on scores) the issue of dimensionality would need to be considered, especially given the different content of the two hands-on exercises.

When super forms are needed, the design moves closer to the currently used NAEP design, in that the items taken by a single examinee do not encompass the full range of content in the test specifications. Even so, the design associated with our proposal to use super forms is much less complex than the current design, so that simpler analysis procedures can be used to produce the results.

When super forms are used, or with a design like the current NAEP design, there are missing data. These missing data require analyses using models, and associated assumptions, that can produce results using the incomplete data. To the extent to which the assumptions of the models are violated, invalid results can be produced. Because of this concern we believe every effort should be made to avoid the use of super forms in order to minimize the number of assumptions needed to produce results.

Estimating Distributions

Presently, NAEP results are reported as properties of latent proficiency distributions for various groups of students. The distribution properties reported are means or the percent of students above certain values on the latent proficiency scale. The concept of latent proficiency distributions can be difficult to explain to policy makers and the public. For this reason, in most major testing programs, normative data are presented in terms of statistics and distributions of observed scores on actual test forms. When interpreting results, test users can be shown a form of the test, and told that the scores reported reflect observed performance on a form like the one they are shown. ACT recommends that in NAEP, as in most major testing programs, scores be reported based on observed score distributions. Smoothing procedures could be used to provide more precise estimates of observed score distributions (Kolen, 1991).

Reporting results using estimated distributions of latent variables is also possible. This section describes procedures for estimating true score distributions using only observed form score distributions, rather than individual item responses.

For the present NAEP design, the proficiency scales are defined based on the latent variables in item response models used for scaling. The latent variables in these models represents the variables measured by the items, where the item responses are assumed to be independent given the latent variables. In the present design for the 1996 NAEP Science Assessment, models are used that assume three latent variables are measured by the items, where each item measures only one of the three latent variables.

To avoid problems with the assumption of local independence at the item level, ACT proposes using latent variable models that model scores at the form level for the purpose of estimating latent variable distributions. Latent variable models could be used with the form score as observed data to estimate a latent variable distribution. In this case there is only one observed variable in the model (the form score) rather than many observed item scores. A general latent variable model for the form score is

$$Pr(X = x_i \mid \alpha,\beta) = \int_a^b Pr(X = x_i \mid \Theta = \theta,\alpha)g(\theta \mid \beta)d\theta \ , \qquad (3.3)$$

where X is the random variable representing the observed form score that takes on the I discrete values $x_1, x_2, \ldots, x_I$, Θ is the random variable representing the latent variable, and α and β are parameter vectors. One class of models of this type are strong true score models (Lord, 1965, 1969). In strong true score models the parametric form of $\Pr(X = x \mid \Theta = \theta, \alpha)$ is given such that $E(X \mid \theta, \alpha) = \theta$, so the latent variable is a true score [defined on an interval (a,b)].

An example of estimating the latent variable distribution is presented using simulated data for the ACT-built NAEP science form. Item response data were simulated with the item response model used in the current design. The item parameters used were those that were estimated in the operational NAEP scaling for the 1996 Science Assessment. The distribution of the three latent variables was taken to be multivariate normal with means all equal to zero and standard deviations all equal to one. The correlations among the three variables were those estimated in the operational 1996 NAEP scaling for science. The three latent variables for each examinee were generated from a multivariate normal distribution. For each simulated examinee, item responses for each item in the form built by ACT were simulated. Two hundred samples of 2000 examinees were generated. The sum of item scores on the 57 items was used as the form score (the minimum form score is 0 and the maximum form score is 70). The true form score is defined as $\sum_{i=1}^{3} \tau_i$ where:

$$\tau_i = \xi_i(\theta_i) = \sum_{j=0}^{n_i} Pr(Y_i = j \mid \theta_i) . \qquad (3.4)$$

In Equation 3.4, θ_i is a value of latent variable i (i = 1, 2, 3), Y_i is the sum of the item scores for the items measuring latent variable i, and n_i is the largest possible score for Y_i. The conditional probabilities on the right side of Equation 3.4 depend on item parameters in the item response models for the items, although this is not explicitly presented in the notation. Note that each of the τ_i is a true score for the sum of the item scores, so that the true form score is not an equally weighted composite of the three true scores on a proportion of total points metric.

The density of the true form score was computed as:

$$h(t) = \int_{L_3}^{U_3} \int_{L_2}^{U_2} g(t - \tau_2 - \tau_3, \tau_2, \tau_3) d\tau_2 d\tau_3 \qquad (3.5)$$

where the limits L_2, U_2, L_3, and U_3 are a function of t determined by the constraints $0 < \tau_i < 1$ and $\tau_1 + \tau_2 + \tau_3 = t$, and

$$g(\tau_1, \tau_2, \tau_3) = \frac{1}{J(\theta_1, \theta_2, \theta_3)} f(\theta_1, \theta_2, \theta_3) , \qquad (3.6)$$

where $f(\theta_1, \theta_2, \theta_3)$ is the multivariate normal density used to simulate the latent variables in the item response model, and the Jacobian $J(\theta_1, \theta_2, \theta_3)$ is given by

$$J(\theta_1, \theta_2, \theta_3) = \begin{vmatrix} \frac{\partial \tau_1}{\partial \theta_1} & 0 & 0 \\ 0 & \frac{\partial \tau_2}{\partial \theta_2} & 0 \\ 0 & 0 & \frac{\partial \tau_3}{\partial \theta_3} \end{vmatrix} = \frac{\partial \tau_1}{\partial \theta_1} \frac{\partial \tau_2}{\partial \theta_2} \frac{\partial \tau_3}{\partial \theta_3} . \qquad (3.7)$$

For each of the 200 simulated data sets, two methods were used to estimate the distribution of the true form scores using only the observed form score distribution. The first method is the four-parameter beta binomial model (Lord, 1965). In the four-parameter beta binomial model the conditional error distribution $\Pr(X = x \mid \Theta = \theta, \alpha)$ is a binomial distribution (in this case there is no α parameter), and the distribution of the true score $g(\theta \mid \beta)$ is assumed to be a four-parameter beta distribution (the vector β has four elements). The specific procedures used to estimate the true form score distribution using the four-parameter beta binomial model are described in Hanson (1991b).

For one of the 200 sets of 2000 simulated examinees, Figure 3.2 gives the observed form score distribution and fitted form score distributions using the four-parameter beta binomial model (labeled

"Observed" and "Beta", respectively). The model appears to provide an adequate fit to the data (the Pearson goodness of fit chi-square statistic is approximately equal to its degrees of freedom). Figure 3.3 gives the true form score distribution as computed using Equation 3.5 with the parameters used to generate the data (labeled "True"). The total score scale is used for the true scores in Figure 3.3 (the true scores range from 0 to 70). Also given in Figure 3.3 is the average of the true form score distributions estimated using the four-parameter beta binomial model over the 200 simulated data sets (labeled "Beta"). The difference between the true and average estimated true form score distributions represents the bias in using the four-parameter beta binomial model.

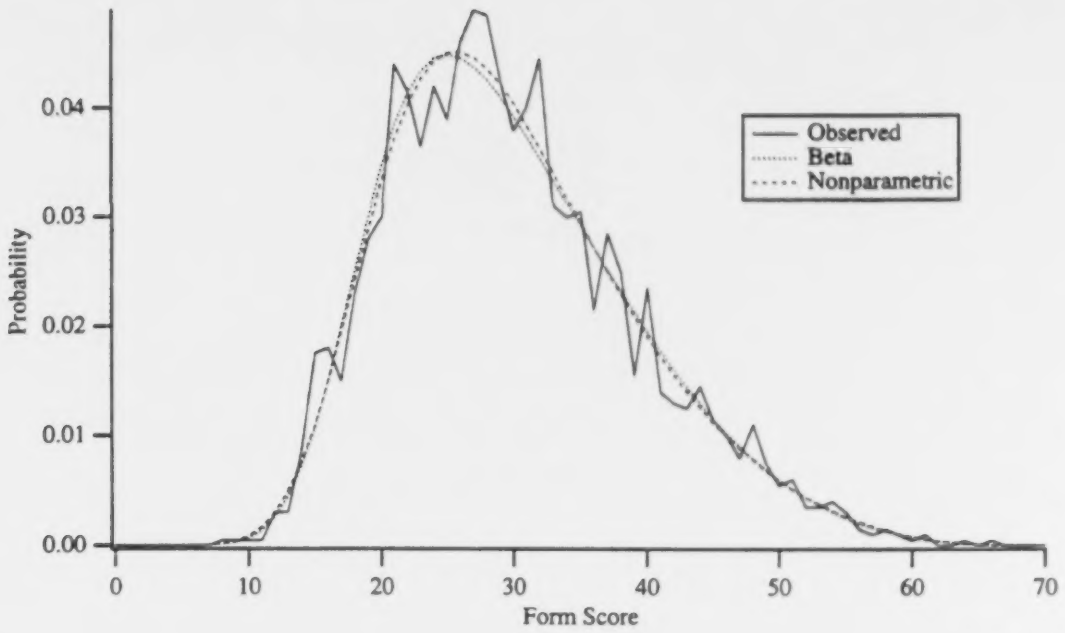


Figure 3.2 Observed and fitted form score distributions for one sample

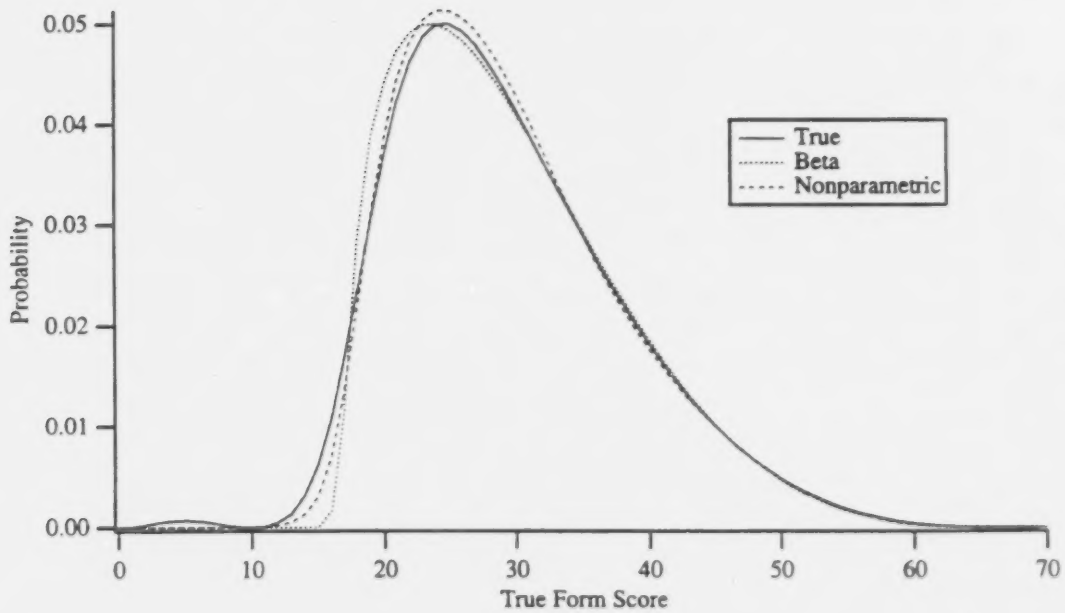


Figure 3.3 True form score distributions averaged over 200 samples

The second method of estimating the true form score distribution maintains the assumption of a binomial error distribution used by the four-parameter beta binomial model, but allows a nonparametric form of the true score distribution (nonparametric in the sense that the true form score distribution is not a parametric function of the true score). A discrete true score distribution is assumed for this method. For this example a discrete true score distribution with 69 values is used. The 69 values on the proportion of possible points scale are given by $j/70$ for $j = 1,..., 69$. The distribution of the observed form score can be written as

$$Pr(X = x_i) = \sum_{j=1}^{69} Pr(X = x_i \mid \Theta = \theta_j)\pi_j, \tag{3.8}$$

where X is the random variable representing the form score (with possible values $x_1 = 0$, $x_2 = 1,..., x_{71} = 70$), Θ is a discrete random variable representing the true score with possible values $\theta_1 = 1/70$, $\theta_2 = 2/70,..., \theta_{69} = 69/70$, and π_j is the probability that the true score is equal to θ_j. The error distribution $Pr(X = x_i \mid \Theta = \theta_j)$ is taken to be a binomial distribution. The probabilities π_j are estimated using a distribution of observed form scores [$Pr(X = x_i)$]. An *ad hoc* two-step procedure was used to estimate the π_j. First, the observed form score distribution was smoothed using the following polynomial loglinear model:

$$\log(m_i) = \alpha_0 + \sum_{r=1}^{6} \alpha_r x_i^{\,r}, \tag{3.9}$$

where m_i is the expected count of the number of examinees obtaining a form score of x_i. The second step involves using the smoothed observed form score distribution as computed using the model in Equation 3.9 to estimate the π_j in Equation 3.8. The distribution of the form score given by Equation 3.8 is a finite mixture. The EM algorithm for estimating the mixing weights of a finite mixture (Dempster, Laird, & Rubin, 1977; Titterington, Smith, & Makov, 1985) was used to estimate the π_j using the smoothed values of $Pr(X = x_i)$.

This nonparametric method was used to estimate the true score distribution of each of the 200 simulated samples of 2000 examinees. The smoothed observed score distribution using the model in

Equation 3.8 is presented for one of the samples in Figure 3.2 (labeled "Nonparametric"). The average of the true score distributions computed over the 200 simulated samples is given in Figure 3.3 (labeled "Nonparametric").

The purpose of this example is to demonstrate how the distribution of the observed form score could be used to estimate the distribution of the true form score. Figure 3.3 shows there is some bias in each of the two methods of estimating the true form score distribution. One source of this bias is the assumption of the conditional distribution of the form score given the true score [$\Pr(X = x \mid \Theta = \theta, \alpha)$ in Equation 3.7] is binomial, which can only be an approximation. Based on the item response model used to simulate the data, the conditional distribution of the form score given the latent variables is actually a sum of multinomial random variables. Given the model used to generate the data, the multinomial random variables are independent conditioned on the latent variables, although in reality they may not be independent (this would be the case if the assumption that item responses are conditionally independent given the specified latent variables does not hold). The binomial distribution should result in an over estimate of the error variance, or an underestimate of the true score variance. A pattern of underestimation of the true score variance is evident in Figure 3.3 (this is especially clear for the nonparametric estimate).

The item response model uses the assumption of conditional independence of the item responses given the latent variables along with an assumption of the parametric forms of the item response functions and latent variable distribution. A latent variable model for the form score does not require the assumption of local independence, although a parametric form of the conditional form score distributions given the latent variable is assumed. The advantage of modeling the form score is that the problematic assumption of conditional independence of item scores is not needed. In addition, item response models are much more complicated to fit and provide much more opportunity for numerical problems to occur. The bias in Figure 3.3 represents bias assuming the item response model currently used is correct. To the extent that the process generating the data is more complicated than can be represented by the item response

model used, the results using the item response model will be biased. It is possible that a simple true score model of form scores, such as those presented using binomial error distributions, could provide less biased results than the item response models that are currently used. It is possible that the effect on the results of the misspecification of models using the form score distributions may be less than the effects on the results of the misspecification of models using item response data.

Besides systematic error (bias) in the estimates of the latent variable distribution it is also important to consider random error. We suspect, for a given sample size, that using latent variable models of form scores will result in less random error in the estimated latent variable distribution than using latent variable models for item responses. The overall error (systematic and random error) might be less for the model using the form score data if the advantage of this model with respect to random error outweighed the advantage of the model using item response data with respect to systematic error (if indeed there were an advantage of the item response model with regard to systematic error). Simulation studies could be used to investigate the issue of the relative random error in estimating the latent distribution using the item response data versus the form score data.

Advantages and Disadvantages of the Proposed Design

The principal difference between the proposed design and the current design is that in the proposed design precise form specifications are used to build parallel forms (or in the case of super forms, nearly-parallel forms) that are completely administered to examinees. A principal characteristic of the proposed design is simplicity. A simpler design has potential advantages and disadvantages relative to a more complex design like the one currently used. The next two sections present some potential advantages and disadvantages of the simpler design that is proposed.

Advantages of Proposed Design

An advantage of a simpler design is that some sources of uncertainty in the results are potentially reduced. The framework presented by Hodges (1987) will be used to present areas where there may be a reduction in the uncertainty in the results.

Hodges (1987) states that statistical activity can be divided for descriptive and analytical purposes into three broad areas: (a) discovery and imposition of structure, (b) assessment of variation conditional on structure, and (c) execution of techniques selected. Each of these three areas has an associated type of uncertainty: structural uncertainty, risk, and technical uncertainty, respectively. The term "uncertainty" refers to doubts about the validity of the results. Invalid results can lead to invalid inferences. The results for NAEP are distributions of proficiency. Invalidities in NAEP results can lead to invalid inferences being made about the proficiency of one or more groups of students. Each of the three types of uncertainty may have potential to be reduced by the simpler design.

Structural uncertainty involves uncertainty in the validity of the statistical models used to obtain the results. A statistical model is invalid to the extent to which the process generating the data is not captured by the model. The term "model misspecification" is also used to refer to invalidity of a model.

In the current NAEP design it is necessary to use statistical models of individual item responses (item response models). One potential type of structural uncertainty in item response models is called "context effects" (Leary & Dorans, 1985; Brennan, 1992). Item response models typically assume that the probability of an examinee responding to an item is not influenced by where the item is located on the test and which items preceed the item on the test. When this assumption is violated, context effects are said to occur. There are several examples where context effects appear to occur with the current NAEP design. An example from the 1994 NAEP reading assessment is given by Zwick (1991) in which invalid results were produced by the use of a item response model which made the assumption that context effects did not occur when in fact they did. Examples of context effects on the 1996 NAEP science assessment are given by Swinton (1997) where item position is shown to have an effect on how often an item is not

responded to. The next section presents some further results using data from the 1996 NAEP Science Assessment for eighth grade that suggest different positions of an item in different NAEP blocks influences performance on the item. Zwick (1991) also indicates that item position had a large impact on percent of examinees reaching certain items for the 1994 NAEP reading assessment.

Other assumptions made by the item response models used in NAEP are possible sources of structural uncertainty. Worthington and Donoghue (1997) examined the effect of the violation of the assumption of local item independence made by most item response models on the validity of the results for the 1996 NAEP science assessment. They found that estimated ability distributions can be grossly distorted if local item dependence (the absence of local item independence) exists for even a single pair of items. They describe changes made to items in the science assessment to attempt to control local item dependence. This is an example of a modification of test content for the purpose of better meeting the assumptions made by the model used for the item data. In the proposed design, this type of modification of test content would not be necessary.

In these examples, sources of structural uncertainty arose due to violations in the models of item response data. In the proposed design, item responses are combined to produce form scores which are the primary data used in analyses to produce assessment results. Some of the sources of structural uncertainty that could possibly exist in item response models would not be a problem in the proposed design where form scores are the data used in the production of assessment results.

The second type of uncertainty presented by Hodges (1987) is risk. Risk, as used by Hodges, involves the uncertainty in the estimation of the statistical models. This is the most familiar type of uncertainty. A large amount of statistical literature is devoted to the assessment and control of this type of uncertainty. Risk can be controlled by the sampling design used to collect the data. In the current NAEP design, background variables are used in the estimation of the reported proficiency distributions in order to reduce risk to an acceptable level. In the proposed design acceptable accuracy of the estimated proficiency distributions should be possible without the use of background variables. The use of the

models which incorporate background variables also could potentially lead to systematic uncertainty due, for example, to the parametric form of the models being misspecified.

The third type of uncertainty is technical uncertainty. Technical uncertainty involves errors that may occur in the execution of techniques used to implement the data analysis. There are two areas of activity in which technical uncertainty can arise: processing of data and application of approximations. Technical uncertainty related to processing of data refers to potential errors in basic data manipulation (e.g., creating data files, merging files, etc.). The simpler design would result in far simpler data processing, reducing the possibility of this type of error.

Application of approximations refers to the approximations used in numerical computations. Examples of these types of approximations include numerical integration and optimization procedures. The models used in the current NAEP design require complicated numerical computations. An advantage of the proposed design is that due to the simpler computations needed it is possible that this type of uncertainty could be reduced. In the proposed design it would be simpler to describe and understand the numerical techniques used. Understanding the numerical procedures is necessary in order to try to assess the extent to which technical uncertainty may affect the results. Details of the specific numerical procedures used in the computations for the current design are not available (undocumented propriety software is used). In the proposed design we recommend that the source code for all numerical procedures be made available so that anyone interested can review the numerical methods used, and could replicate the results.

Disadvantages of Proposed Design

A major disadvantage of the proposed design relative to the current design is that the construction of test forms with tight specifications administered in their entirety to examinees results in less flexibility in choosing items, since the items to be included on a form would have to meet the tight form specifications. On the other hand, the item response models used in the current design can result in

restrictions on items due to the need for items not to violate the assumptions made in the models (Worthington & Donoghue, 1997).

In a design like the current design where only a small portion of the total items are presented to an individual examinee there is a potential for shorter testing times than in the proposed design were examinees take entire test forms (although the proposed design does not require any additional testing time over that used in the current design). Shorter testing times may potentially alleviate negative effects on the results due to low motivation of test takers to do well on the test.

Summary of Advantages and Disadvantages of the Proposed Design

The proposed design has advantages and disadvantages relative to the current design. The framework of Hodges (1987) was used to present some possible advantages of the proposed design in terms of reducing the amount of uncertainty in the results. All three types of uncertainty are present in both designs. We think a strong case can be made that the simplicity of the proposed design can result in less uncertainty. We have presented evidence of some sources of structural uncertainty in the current design that have produced invalid results, or have the potential to produce invalid results.

We have built one form of the eighth grade science assessment that met the Framework for the science assessment to a reasonable degree. We believe test forms that are completely administered to individual examinees and constructed to precise specifications have the potential to provide coverage of an adequate range of content for the NAEP Science Assessment. The current design offers more flexibility in content coverage, but with potential for more uncertainty in the results. There is an inevitable tradeoff between flexibility of test construction and uncertainty of results. We believe that adequate content coverage can be attained with a much simpler design. Whether the advantages of the simpler design outweigh the greater constraints on test developers that would result is a judgement to be made by the policy makers for NAEP.

An Investigation of Context Effects Based on Item-Block Position on the 1996 NAEP Science for Grade 8

In the present NAEP BIB design, blocks of items change positions. The scaling model assumes that the position has no effect on the item characteristics. In this section, we examine this assumption. The violation of this assumption could increase structural uncertainty in the results, as described in the previous section.

Assessment items in the 1996 NAEP in Science are administered in blocks. There were three types of blocks: Concept/Problem Solving (CP), Theme-Based (TB), and Hands-On (HO). The first two types are paper-and-pencil blocks with both multiple-choice (MC) and constructed-response (CR) items CP blocks include items from the three fields of science identified in the NAEP Science Assessment framework. TB blocks include items related to one theme in the framework, and most of these blocks include items from one single content area. HO blocks involve performing a task and responding to items related to the task or the results of the experiment performed to complete the task. For each grade level assessed by NAEP, there were eight CP blocks, three TB blocks, and four HO blocks, for a total of 15 blocks.

Each form of the 1996 NAEP Science Assessment contained three blocks of test items -- two paper-and-pencil blocks and one HO block. Forms were constructed with one HO block plus either two CP blocks or one TB and one CP block. This scheme of combining blocks for assessment forms would yield 208 possible forms. Only 37 forms were used in the assessment, however. The HO block was always in the third section of each form.

This investigation of context effects focuses on the main effects of block position on student performance on individual items and on the percentages of nonresponse to those items. Since only paper-and-pencil blocks change position, HO blocks were not included in this investigation. Moreover, all the analyses were performed on grade 8 data only.

Background Studies investigating effects of rearranging sections of items in a test were reported as early as the 1950's (Leary & Dorans, 1985). The question under investigation was whether an item appeared to be easier or more difficult when it appeared earlier in a test than when it appeared later.

Mollenkopf (1950) studied the effects of rearranging sections of test items on item difficulty and item discrimination for verbal and mathematics aptitude tests. Under power conditions, Mollenkopf found that for the verbal test an item is easier when the item appeared early in the test (i.e., first section) than when it appeared late in the test (i.e., third section), but no similar position effect was found in the mathematics test. Under highly speeded conditions, however, both item difficulty and item discrimination were affected by position on both the mathematics and verbal tests. Items were found to be more difficult and had higher discrimination indices when they came late than when the items came early in the test.

Brennan (1992) reported that in a 1988 study of the original ACT Assessment, item position did affect examinee performance. The test form used for this study had four passages with associated test items and a set of discrete items. A scrambled form was created by interchanging the positions of the first two passages and the last two passages. There was no change in item order otherwise. Results indicate that items related to the first two passages and the discrete items were unaffected by the rearrangement. However, items in the fourth passage were more difficult in the original form where it was the last passage than they were in the scrambled form, and items in the third passage were more difficult in the scrambled form, where it was the last passage than they were in the original form. A hypothesis that was cited for this context effect is that fatigue causes examinee performance to deteriorate at the end of the test.

Block position effects on the 1996 NAEP Science Assessment was investigated by Swinton (1997). The focus of his inquiry was mean group performance across the two block positions rather than characteristics of individual items. He reported significant differences in group performance for grades 4 and 8, although the differences are in the opposite directions. In grade 4, position 2 group performed better than position 1 group, and the reverse is true for grade 8. There were no significant differences across group performance for grade 12.

Analysis and Results

More than 10,000 students took the grade 8 science assessment. Each student took three blocks of items, the first two of which were paper-and-pencil blocks. About 2,000 students took each of eleven CP or TB blocks. Each student who took each block was assigned to one of two groups based on whether that block was in the first or second section of the form that the student took. For each block the number of students in each group are about equal. The item difficulty, percentages of nonresponse (i.e., "omit" or "not-reached") were compared for the two groups for each item. The three TB blocks contain 12, 10, and 13 items, with two, five, and three MC items, respectively. Each CP block has 16 items with seven or eight MC items each.

The percent correct or p-value of each item was used as the index of item difficulty. For MC items, the p-value was computed as the percent of students who selected the correct response. Omits and multiple responses were considered incorrect, and examinees who did not reach the items were not included in the computation.

For each CR items scored for partial credit with n score levels 0, 1,..., n-1, where 0 indicates an incorrect response and n-1 indicates a correct or complete response, the p-value is the average score divided by n-1 and the quotient is multiplied by 100. Omits, "off-task," and "not-rateable" were considered incorrect. "Not-reached" were not included in the computation.

Results of comparisons of p-values, and percentages of omits and not-reached are presented in Figures 3.4 through 3.9. Figures 3.4 through 3.6 compare the three item statistics by item format (MC versus CR), and Figures 3.7 through 3.9 present the comparisons by type of block (TB versus CP). Results of analyses of variance are in Table 3.2.

Table 3.2 Comparisons of Average Overall P-Values, and Percentages of Not-Reached and Omits

Items	Overall P-Value			% Not-Reached			% Omits		
	First	Second	Level of Significance of Block Position Effects	First	Second	Level of Significance of Block Position Effects	First	Second	Level of Significance of Block Position Effects
All (n=163)	**39.32**	**37.98**	0.0001	2.11	2.51	0.0001	**2.97**	**3.58**	0.0001
MC (n=71)	**52.82**	**51.93**	0.0027	0.53	0.67	0.0001	**0.39**	**0.51**	0.0121
CR (n=92)	**28.90**	**27.22**	0.0001	3.333	3.93	0.0001	**4.96**	**5.96**	0.0001
Level of Significance of Item Type Effects	0.0001	0.0001	**0.0175**	0.0001	0.0001	**0.0002**	0.0001	0.0001	**0.0001**
TB (n=31)	**48.12**	**46.19**	0.001	1.24	1.39	0.0014	**3.69**	**4.86**	0.0001
CP (n=132)	**37.25**	**36.05**	0.0001	2.32	2.78	0.0001	**2.80**	**3.28**	0.0001
Level of Significance of Block Type Effects	0.0116	0.0175	**0.0789**	0.1878	0.1151	**0.0453**	0.2850	0.0799	**0.0041**

Notes: 1. The averages of % omit, % not-reached, and overall p-values are in **bold**.
2. In shaded cells are the levels of significance of interaction effects.

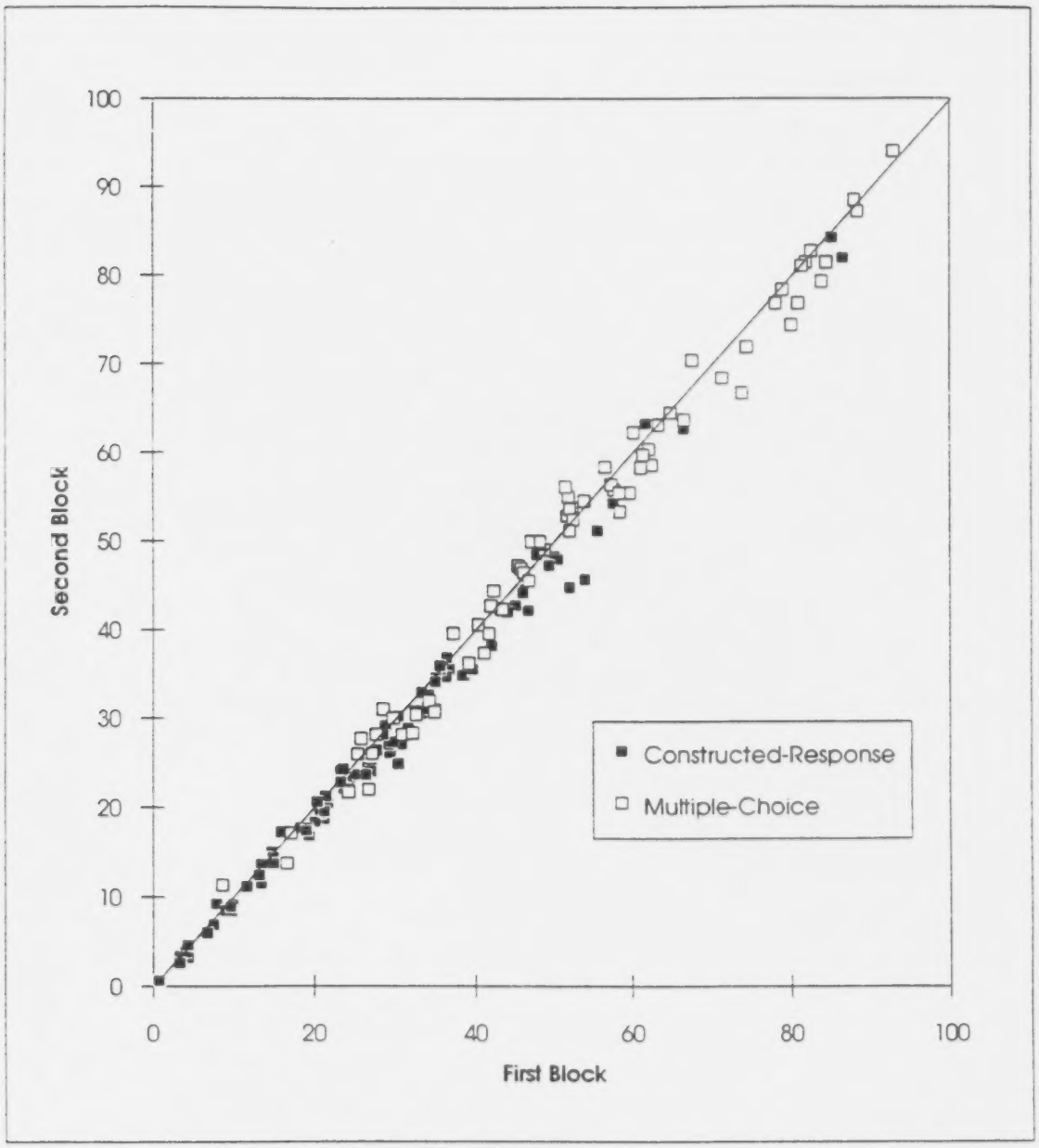


Figure 3.4 Overall P-Values, By Item Format

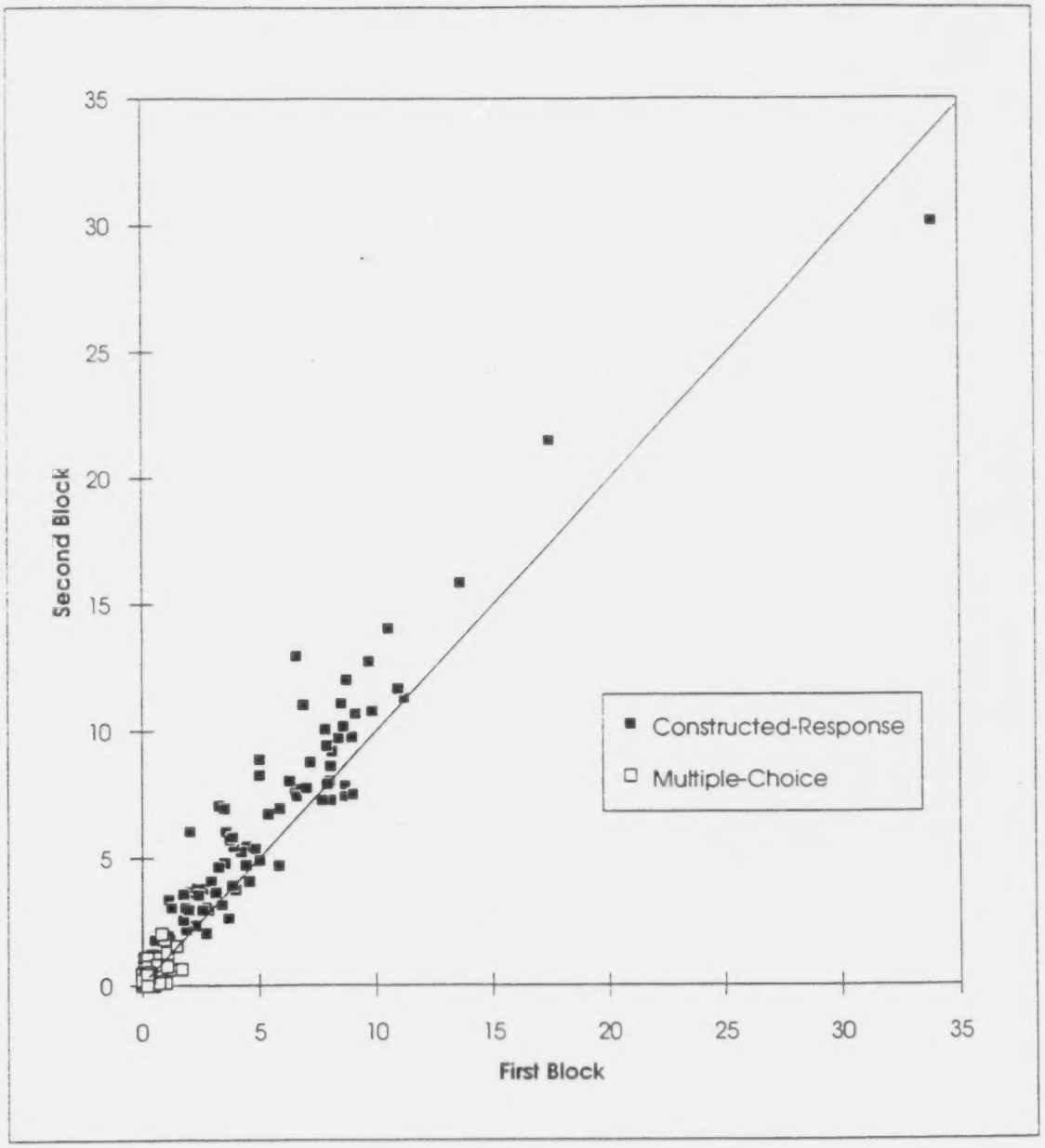


Figure 3.5 Percentages of Omits, By Item Format

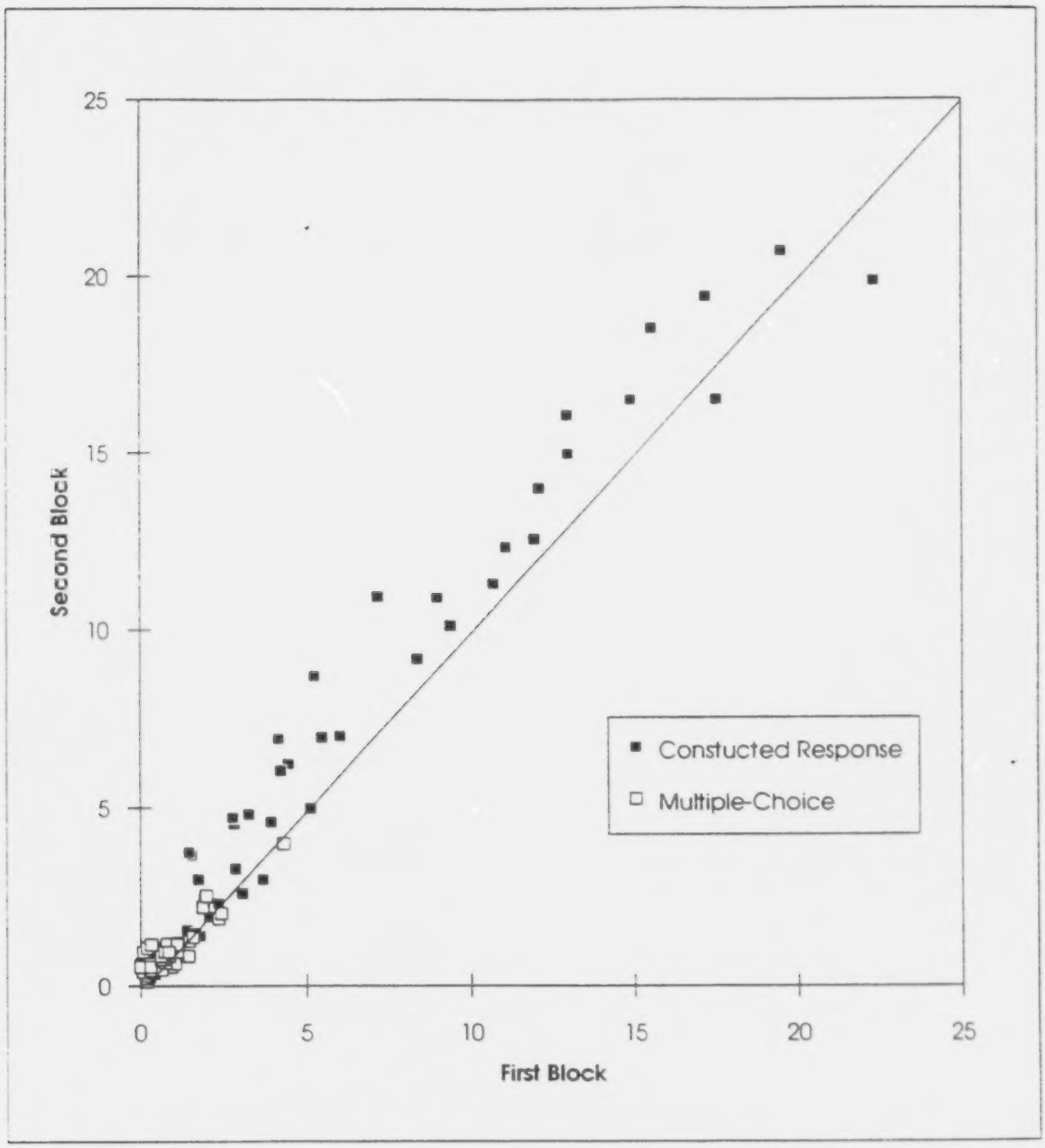


Figure 3.6 Percentages of "Not-Reached," By Item Format

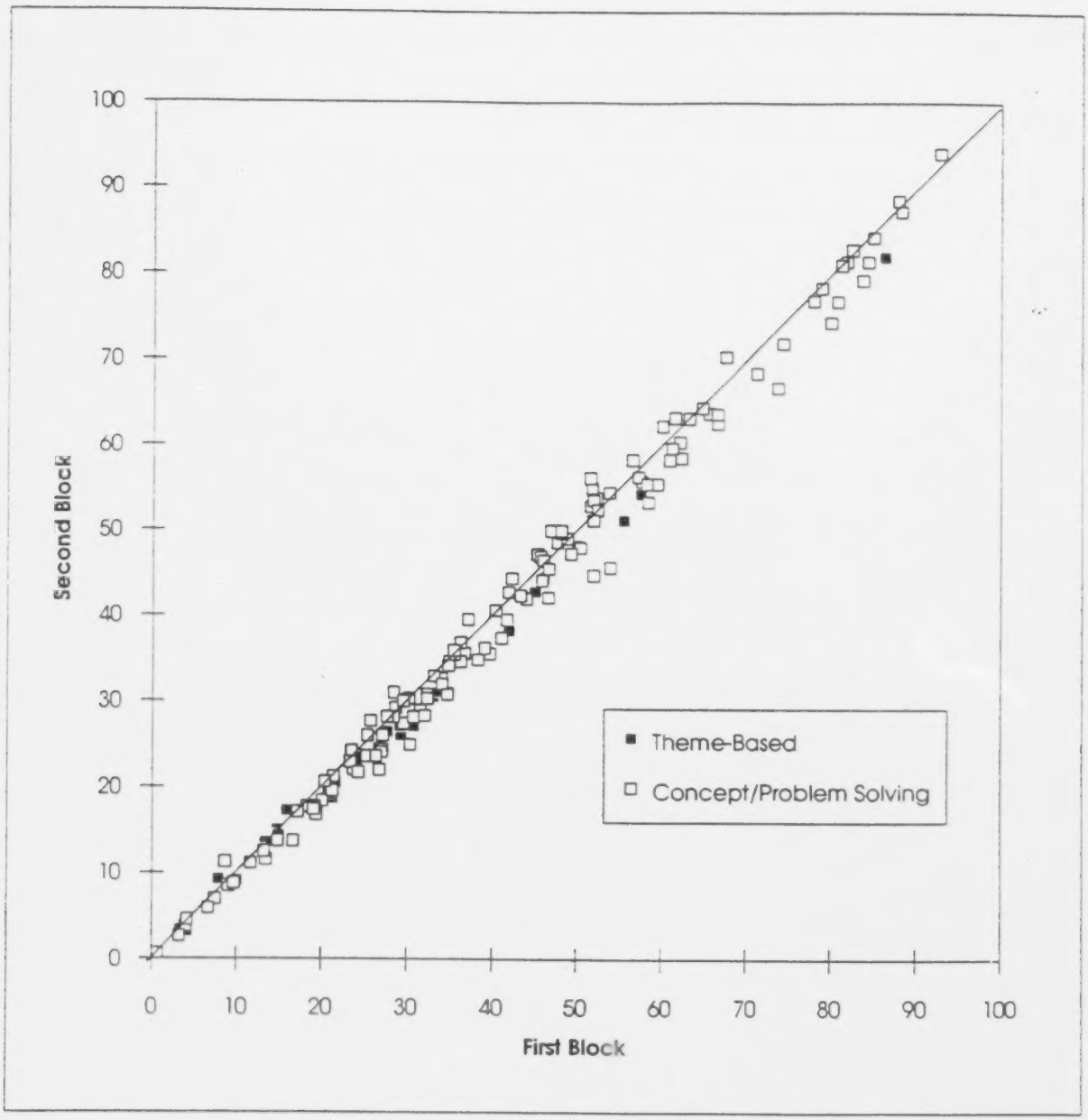


Figure 3.7 Overall P-Values, By Block Type

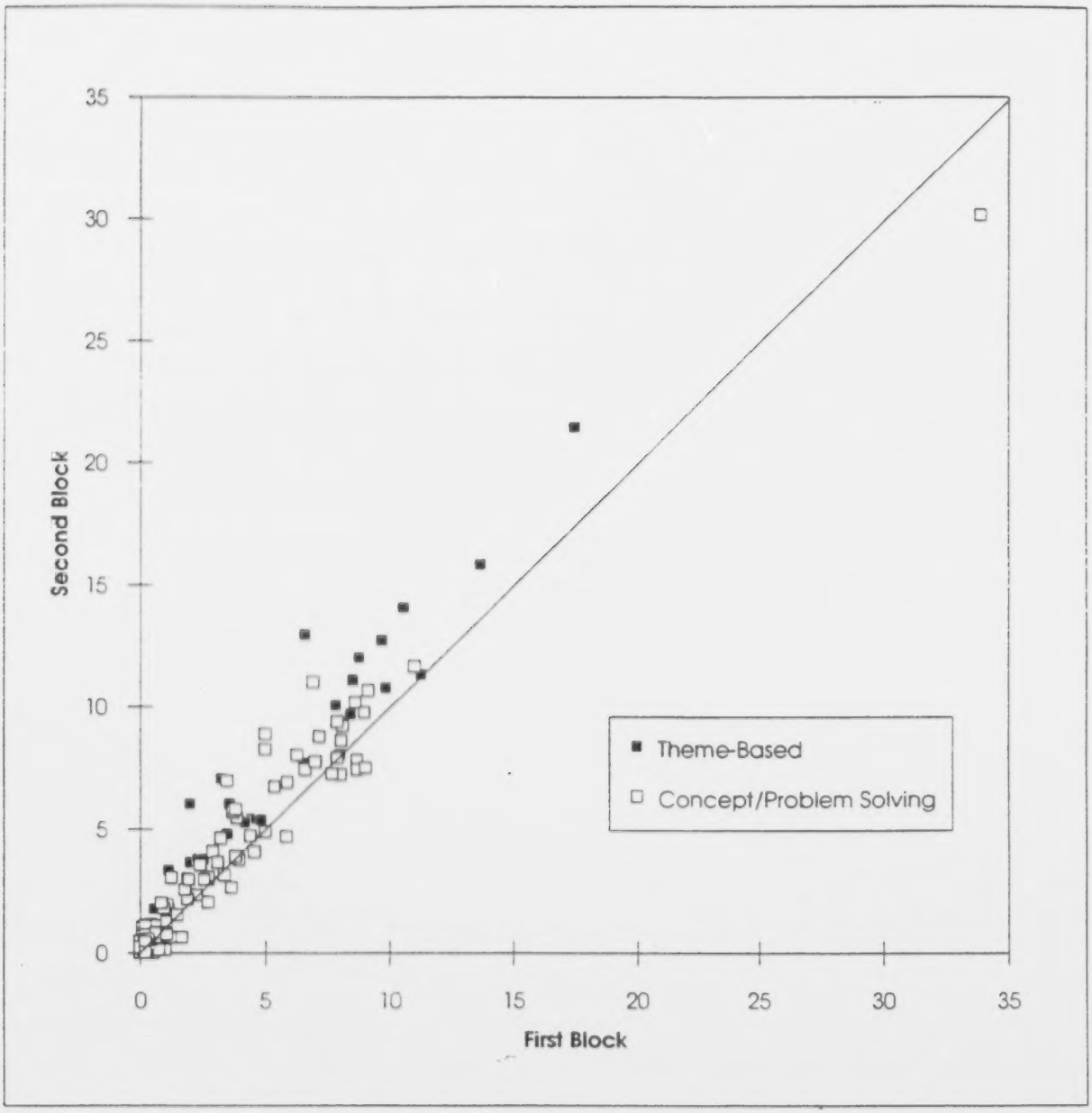


Figure 3.8 Percentages of Omits, by Block Type

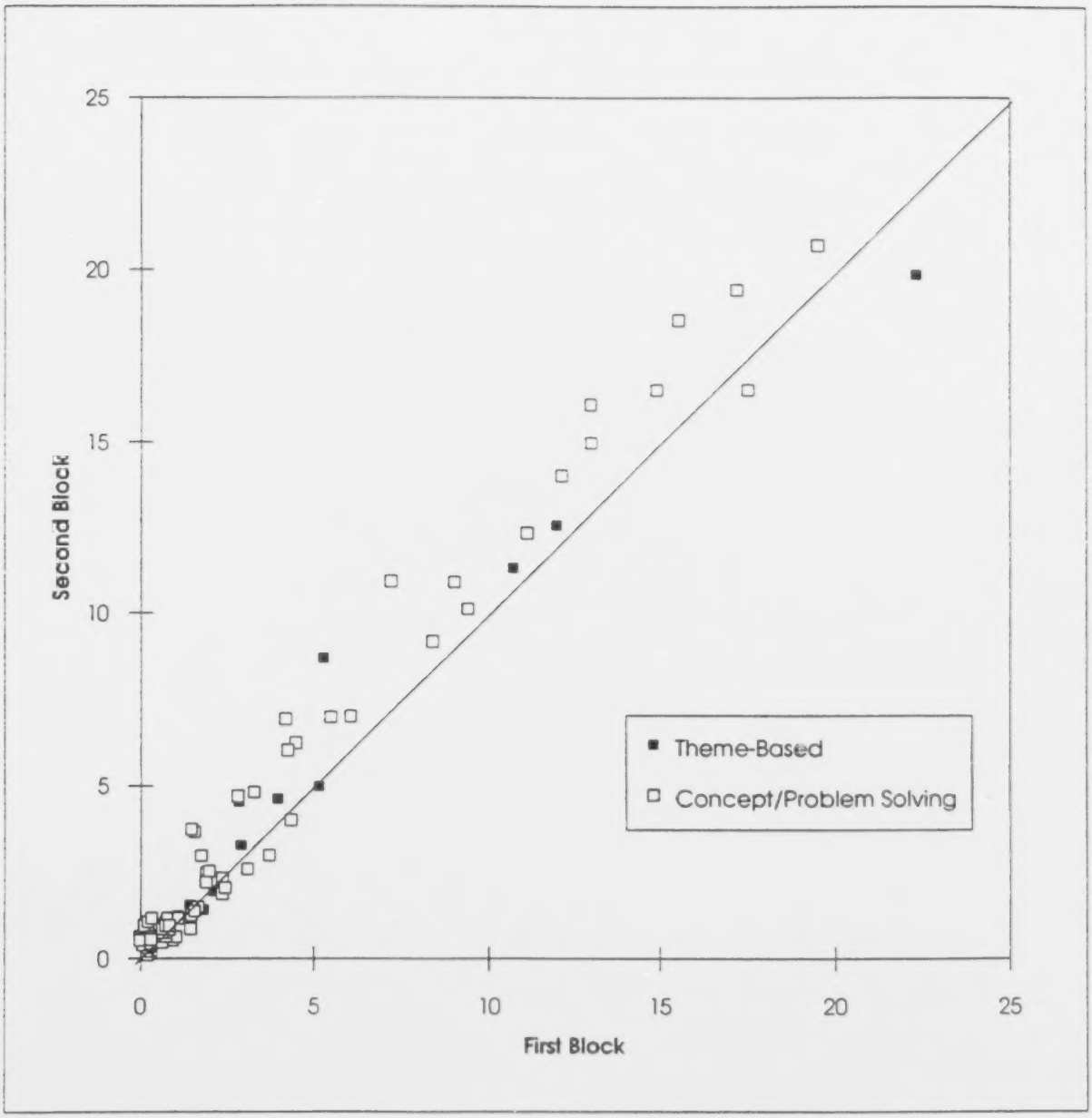


Figure 3.9 Percentages of "Not-Reached," By Block Type

Each point below the diagonal lines in Figures 3.4 and 3.7 represents an item that has a p-value that was higher when the item appeared in the first section of the test form than when it appeared in the second section. Items seem to be more difficult when they were in the second section of the test form than when they were in the first section. Moreover, MC items appear to be easier than CR items, as seen in Figure 3.4. Figure 3.7 does not seem to indicate difference in item difficulty by block type.

Figures 3.5 and 3.8 show that items generally have higher percentages of omits when they appear in section 2 of the test than when they appear in section 1. Clearly, examinees tend to skip items in the second section of the test more than they do in the first section. The percentages of omits, however, are significantly higher for CR items than for MC items. There were no significant differences in percentages of omits by block type.

The percentages of "not-reached" are generally higher when items appear in section 2 than when they appear in section 1 as seen in Figures 3.6 and 3.9. Percentages of "not-reached" are definitely higher for CR than for MC items as shown in Figure 3.6. This occurrence, however, might be due to test construction more than the students' test-taking behavior. Except for one TB and one CP blocks, the last three items in each block are CR items.

Differences in item difficulty and nonresponse rates seem to indicate that the test items in each block might not be presenting the same cognitive tasks to all examinees, that the difference in the position of the block alters the cognitive demand of each item. If this hypothesis is true, and the items function differently when their position in the test form are altered, then the assumption of item parameter invariance in IRT, which is used in NAEP scaling, might have been violated.

Summary

The following items summarize the crucial elements of the proposed psychometric design.

1. Write detailed specifications for the test forms. This is the most crucial aspect and the centerpiece of the proposed design.

2. Determine whether forms administered in full to examinees (administered forms) can be developed that completely reflect the specifications. If not, construct administered forms in which the unique items across administered forms (a super form) meets the specifications. If super forms are needed, create the administered forms that constitute a super form to have as many common items as possible, where the common items are as near to meeting the specifications as possible.

3. Define a score on a form or super form that will be used in the analyses to produce results. Analyses will be based on estimated distributions of this form score for populations of interest. In the case that super forms are needed, application of missing data methodology or latent variable models can be used to estimate these distributions.

4. If it is desired to report results using latent variable distributions, estimate these from the observed form score distributions computed in step 3.

A major distinction between the proposed design and the design currently used is that the proposed design puts more of the burden for the validity of the results on the specification and construction of examinations and less of the burden on complex analytic procedures.

CHAPTER FOUR
SAMPLING METHODOLOGY

In this chapter, we present some simplifications to NAEP methodology, the rationale for these simplifications, and a simulation study to compare present NAEP sampling methodology to the proposed methodology. The results of the simulation show that the simplifications would reduce the number of schools required to participate in NAEP testing. In addition, we make a suggestion on how to combine the state and national samples in a way that allows for the use of schools in the state assessment for the national assessment.

Sampling Methodology

The current NAEP sampling methodology utilizes a four-stage cluster sample. This method is a fairly straightforward and commonly used design; ACT would propose that the basic design remain the same, but with some simplifications. The current method defines the primary sampling unit (PSU) as a geographic unit such as a Metropolitan Statistical Area (MSA) as defined by the U.S. Bureau of Census. Some of the PSUs are selected with certainty. The remainder are selected within strata defined by various socioeconomic and geographic variables. Selection of schools within PSUs is the next stage, followed by selection of session type and finally selection of students within schools. At each stage, where appropriate, selection is made with probability proportional to size, or a measure of size.

Proposed Changes to Sampling Methodology

We propose that the primary sampling unit be designated as the school, which would eliminate the first step in the current process. Choosing schools as the primary sampling unit is the current methodology used for the state NAEP assessment. In addition to simplifying the procedure, use of schools as PSUs allows consideration of combining state and national assessment samples, as suggested below. Although not addressed here, one should study the effectiveness of alternative definitions of strata.

Geographic representation can be achieved by dividing the nation into an appropriate number of regions and selecting schools within each region. Stratifying by geographic region is a common technique and is similar to what is currently being used. Also, strata will be defined by the size of the school, to control for bias in the resulting estimates, as discussed below.

In addition, we propose that the schools be selected with equal probability within strata, rather than proportional to size, and within each school, all eligible students be tested. Sampling proportional to size is an effective methodology when the cluster size and the cluster total are highly correlated [see Cochran (1977), p. 295]. The corresponding quantities here are school size and school ability. These variables are not highly correlated, and so selection proportional to size does not yield large gains in efficiency. Sampling with equal probability for each school and testing all eligible students makes the sample an equal probability selection method, eliminating some of the need for complex weighting of cases. We believe that the simplified design is a more efficient use of resources and the same, or perhaps better, precision can be achieved with reduced cost.

When all students within a school are tested, the sample size is also a random variable, since we do not know which schools are chosen in advance. Random sample size can introduce a bias into ratio estimates of means and proportions. To diminish this bias, we need to maintain some control over the sample size, so that its variability from sample to sample is not large. To maintain this control, we use size of school as a stratification variable. We divide the schools into strata, based on their size, and sample from each stratum proportionally. Dividing schools in this way guarantees that we have a certain number of small, medium-sized and larger schools in the sample. The larger the number of strata, the lower the variability of the sample size, and consequently, the lower the bias. A coefficient of variation of sample size is calculated for each sample; so long as this coefficient is below .1, the bias will be small [see Kish (1965), pg 208].

Rationale for Proposed Changes

The proposed methodology has advantages and disadvantages compared to the current method. By eliminating the sampling of MSA's, the proposed method is simpler. The simplicity is further enhanced by testing all grade-eligible students, rather than using a selection method that relies on field staff to make appropriate adjustments regarding sample size and requires the school to make a roster available well ahead of time. We have found in our sampling procedures, that schools would prefer to test all students rather than be forced to sample a selected few from out of their regular classes and make special accommodations for testing. Very large schools may prefer to test only a part of their class, for logistical reasons, and this can be accommodated by allowing subsampling within a school in the large school stratum.

Sampling schools directly rather than from MSA's will also likely lead to a more efficient sample; i.e., a smaller sample size would be needed to achieve a given precision target. The reason is that schools drawn randomly from within an MSA are likely to be more similar than two schools drawn randomly from a larger grouping (e.g., stratum) of schools. The greater the similarity of schools within an MSA results in a greater redundancy of information.

On the other hand, sampling all students within a school, rather than selecting a specific number to be sampled within a school, will lead to a larger sample size for a fixed precision level. The reason is that two students randomly selected within a school are likely to be more similar than two students randomly selected from different schools. From an administrative point of view, the proposed method is more efficient in that fewer schools are required. It is easier to sample 50 schools with 100 students each rather than 100 schools with 50 students each. The choice involves some compromise between the two efficiencies. To get an idea of the size of the difference, we have done a small simulation study comparing design effects of the two methods. The design effect (DEFF) for estimating a parameter using a particular sample design is the ratio of the sampling variance of the parameter estimate to that of the sampling variance of an estimate based on a simple random sample.

Simulation to Compare Methodology

In our simulation study, ACT scores were used in place of NAEP data, due to the lack of access to the latter at the time we conducted the analyses. We considered data from the ACT Assessment history file for June 1996, taking precautions to prevent any duplicates. The number of students from a school who tested on that date is considered as the "population" from that school. To obtain reasonable sizes for the schools, the states from which the data were analyzed were limited to the 28 states where the ACT is the dominant test.

For the study, 100 replications are used. In each replication, 200 schools were chosen and each student within the school was included. The variance was calculated for the stratified cluster sample, using the usual variance formulas for ratio estimators [see Cochran (1977), Chapter Six]. The median and range for coefficient of variation and the standard error for the estimated proportion of students with Composite ACT scores of 20 or less. This value was chosen to be as close to the median Composite score as possible.

For the current methodology, we also used a Monte Carlo approach and tried to stay as true to the sampling procedure as was possible. We first split the sample into geographic PSUs. We used Consolidated Metropolitan Statistical Areas and Metropolitan Statistical Areas and finally counties for those schools which are not in either of the above categories (counties were grouped in cases where the size in a county was less than 250 tested students). Some of the regions were included with certainty and the rest were chosen with probability proportional to size. The PSUs included with certainty were chosen to mimic the ones chosen for the National NAEP sample in 1994. Noncertainty PSU's were chosen with probability proportional to size. In this case, we used as a measure of size, the number of tested students, rather than the population of the geographic region. Within each chosen PSU, 4 schools were chosen, if possible. Within each school, a random sample of 40 students was then selected. If there were fewer than 40 students at that school with an ACT score, then all the scores were included. Weights were calculated to adjust for the unequal chance of selection. The variance of the proportion of students with scores at or

below 20 was calculated. The variance was calculated using the same formula as above. In the NAEP project, variance is calculated using a jackknife procedure, but, for the sake of comparison, this calculation was not used here. As before, we repeated this procedure 100 times and report the median and range for the standard error of the percentage below 20.

Results

Table 4.1 below gives the median values of the design effect for the 100 samples. In addition, it lists the minimum values and the maximum values. The last column gives the rate of homogeneity associated with the median design effect.

Table 4.1. School size, median, minimum and maximum design effects and median rate of homogeneity for the two sampling plans

	Average School Size	Design Effect median min max	Rate of Homogeneity
Sampling with school as the PSU	61.2	5.07 4.11 6.40	.071
Sampling with region as the PSU	30.5	3.42 2.69 4.15	.082

The median design effect for the current sampling methodology is smaller than with the proposed methodology, but this is due solely to the fact that the average school size is smaller. The rate of homogeneity is smaller with the proposed methodology. This gain represents the effect of the similarity within the PSU for the current methodology.

The result of this study suggest that changing to the new methodology will lead to sample sizes that are slightly larger in terms of the number of students, while being slightly smaller in terms of the number of schools, typically about 15%. For example, if, using the current methodology, 1,000 students are tested with a single form, then the approximate equivalent sample size required to achieve the same precision with a simple random sample is about 133. This assumes about 80 tested students per school, in line with what was achieved in 1994 main samples. This would require about 13 schools. Using the

proposed methodology, it would take about 1068 students and 11 schools to reach the same precision. This assumes an average of 100 students per 8th grade, a figure calculated using a database of schools from Market Data Retrieval. The number of students will only be slightly larger since the current methodology tests all students in a school except in the largest schools.

In general, the proposed method will require fewer schools and more students. This is due to the choice made to test all students per grade, in addition to the slightly more efficient design.

State and National Assessments

One of the concerns regarding the current NAEP sampling design is the duplication of effort required to test at both the national and state levels. It seems, for example, that a school in California used in the national sample could also be used in the California state sample. Currently, that is not the case. Due to differences in the sampling methodology and the population of interest, the two samples are drawn separately. Some of the changes in the NAEP (e.g. the elimination of age-eligible restriction) combined with some of the suggestions that we have made previously allow for the samples to be combined in a simple fashion.

Assume, for the sake of illustration, that there is only one assessment for both the state and national level. The dual state-national sample could be collected in the following manner. Within each state, choose a sample of schools for the state assessment. For each state, we would identify a subsample that will be identified as the "national sample". The number of schools so identified would depend on the size of the state.

If a state decided not to participate in the NAEP assessment for that particular year, the schools in the "national sample" would still be contacted for the national assessment. The rest of the schools in the state would not be contacted.

As an example, suppose there are only two states, one of which is three times as large as the other. Suppose that both the state and national samples require 100 schools. We would select 100 schools from

each state for the state assessment. In the larger state, 75 of these schools would be identified as part of the "national sample". In the smaller state, 25 schools would be so identified. If both states participated, all 200 schools would be contacted for the assessment. The data from all schools would be used in the national assessment. The data from the smaller state would be weighted by 1/3, to account for their overrepresentation in the sample.

Note that this is only feasible if sampling is done at the school level. Starting with a PSU of an MSA which are chosen with differential probabilities would make this method very complex. It might also be necessary to combine some of the smaller states for identification of schools in the "national sample". Still, this method is simple and would allow for fewer schools to be assessed than the current method.

Discussion

We believe that our proposed sampling methodology is more straightforward and more efficient than the present methodology. The rate of homogeneity is reduced about 15% by using school as the PSU. The gain is due to the elimination of the first stage of sampling in the current methodology. In addition, sampling all students rather than a portion will reduce the number of schools required to test. This gain will be smaller, as the current practice is to test all students except in the largest of schools. It will eliminate some of the complex weighting that is required with the current method.

Sampling at the school level also allows for introduction of a simple method to combine the state and national assessments. This method eliminates the need for separate samples to be drawn for both the state and national assessments.

CHAPTER FIVE
FURTHER PSYCHOMETRIC ANALYSES

Projecting to a Domain

If a domain score or marketbasket approach is to be adopted for NAEP, it will be necessary to develop techniques to estimate performance for the group in the domain or marketbasket, given the various individual forms taken. A domain in our terminology would be equivalent to a group of super forms. Under ACT's proposed NAEP design, an individual would take one form, which may or may not be treated as a subset of the domain when estimating domain performance. Because an individual would not take the entire domain, it is necessary to determine whether domain performance can be accurately estimated from available information: namely, performance on the form taken and the characteristics of the items in the domain.

Research is currently being performed at ACT to evaluate methods for estimating domain scores from an individual form (e.g., Pommerich & Nicewander, 1996, 1997). The domain score research at ACT was undertaken to evaluate the feasibility of estimating school-level IRT-based content domain scores under conditions where there are small numbers of items per content area on a test form, small domains, and small numbers of observations per school; conditions which may be applicable under NAEP's sampling rubric. A promising method has been developed in which an estimated distribution of school ability is computed based on the individual form taken, and the estimated domain score is computed as a function of the estimated distribution and the item characteristic curves for items that comprise the domain.

ACT research suggests that school-level domain scores computed in this manner are more accurate estimates of actual domain performance and more consistent over individual forms taken than the observed performance on the individual form taken, particularly when the individual forms vary in terms of difficulty. However, the research was performed to evaluate domain score estimates under conditions specific to ACT's tests. To determine the suitability of this domain score methodology for NAEP, we

conducted a simulation study where item responses to two NAEP test forms were generated using NAEP Science parameters. One form was ACT's prototype NAEP Science test form described in Chapter Two; the other was the operational NAEP Science test form S208D (blocks F, H, and J). Item responses were generated for the items on each form. For each form, the remainder of the items in the Science item pool was treated as the domain; item responses were also generated for these items. Methodology employed in previous studies (Pommerich & Nicewander, 1996, 1997) was adapted to include both multiple choice and open-ended items with varying numbers of response categories. Group-level domain score estimates were then compared to the actual group domain scores.

Domain Score Definitions

A domain score for an examinee, where the domain consists only of multiple choice items, can be defined as the percentage of items in the domain the examinee can answer correctly. A domain score for a group, then, is the average percent correct in the domain for students within the group.

When the domain consists of both open-ended items that are scored polytomously and multiple choice items, a domain score for an examinee can be defined as the proportion of total possible domain points received by the examinee. The domain score for the group would be the average proportion of points for students within the group.

Item response theory (IRT) provides a convenient method for estimating domain scores from performance on the form taken, and known domain item parameters. Using this approach, a domain score for an individual, where the domain consists only of multiple choice items, may be computed as

$$\frac{1}{J}\sum_{j=1}^{J} P_j(\theta), \tag{5.1}$$

where θ is an IRT scale score of examinee ability and $P_j(\theta)$ is the probability of answering domain item j correctly at ability θ. In practice, θ is unknown, and an estimate of examinee ability is employed instead. A group domain score may be computed as the average of student domain scores.

One benefit of a domain score defined in this manner is that θ can be estimated from one set of items (i.e., the form taken), while the domain score can be determined from a set of items not taken by the examinee (i.e., the content domain), so long as the item parameter values for the items not taken are known (Hambleton & Swaminathan, 1985). Note that under ACT's proposed design, the domain would consist of a group of individual forms or super forms rather than a pool of individual items. Assuming the domain item parameters are known and on the same scale as parameters for the items taken, the prevailing problem is to accurately estimate ability—a problem that is exacerbated when the number of items taken is small. Group-level domain scores may also be adversely affected by small domains and small numbers of students per group.

For this project, two different methods of estimating group domain scores were employed. In the first method, a point estimate of ability was computed for each student within a group based on item responses to the single form taken, and the estimated domain score for the group was computed as

$$\frac{1}{T}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=0}^{K_j-1} kP_{jk}(\theta_i)\,, \tag{5.2}$$

where $P_{jk}(\theta_i)$ was the probability of student i on domain item j being in score category k (with K_j possible response categories), and T was the total number of points possible on the domain items (i.e., $\Sigma_j\,(K_j - 1)$). $P_{jk}(\theta_i)$ was computed from a three-parameter (3PL) logistic model for multiple choice items ($k = 0,1$), or from a two-parameter (2PL) logistic model for open-ended items scored with only two response categories ($k = 0,1$). For open-ended items with three or more response categories ($k = 0,1,2,...,K\text{-}1$), $P_{jk}(\theta_i)$ was computed from a generalized partial credit model (GPCM; Muraki, 1992). Student-level ability was estimated from the mean of the posterior distribution (*expected a posteriori* estimate; Bock & Mislevy, 1982), based on item responses to the form taken. Domain scores estimated using this method are labeled EAP.

In the second method, a distribution of group ability was computed based on examinee responses to t items on the single form taken, and the estimated domain score over J domain items was computed as

$$\frac{1}{T}\sum_{j=1}^{J}\sum_{k=0}^{K_j-1}\int_{-\infty}^{+\infty} kP_j(\theta)f_g(\theta)d\theta \equiv \frac{1}{T}\sum_{j=1}^{J}\sum_{q=1}^{m}\sum_{k=0}^{K_j-1} kP_{jk}(\theta_q)p_g(\theta_q)\,, \tag{5.3}$$

where $P_{jk}(\theta_q)$ was the probability of being in score category k on domain item j given ability θ_q, and $p_g(\theta_q)$ was group g's probability that ability = θ_q for $q = 1,2,...,m$ theta points. Maximum likelihood estimates of the probabilities $p_g(\theta_q)$ were computed using the EM algorithm. [For notational convenience, $p_g(\theta_q) = \pi_q$.]

The E-step at iteration s consisted of computing for each q (see Woodruff & Hanson, 1996)

$$n_q^{(s)} = \sum_{i=1}^{N} \frac{\left[\prod_{r=1}^{t}(P_r(\theta_q)|y_{ir})\right]\pi_q^{(s)}}{\sum_{l=1}^{m}\left[\prod_{r=1}^{t}(P_r(\theta_l)|y_{ir})\right]\pi_l^{(s)}}\,, \tag{5.4}$$

where $n_q^{(s)}$ was the number of the N examinees within group g for whom ability is equal to θ_q, y_{ir} was the score of examinee i on item r on the form taken ($y_{ir} = 0,...,K_j-1$). Normalized densities from a normal distribution were used as initial starting values for the $\pi_q^{(0)}$.

The M-step at iteration s consisted of using the $n_q^{(s)}$ computed in the E-step to compute

$$\pi_q^{(s+1)} = \frac{n_q^{(s)}}{N}\,. \tag{5.5}$$

The values of $\pi_q^{(s+1)}$ computed in the M-step at iteration s were used in the E-step at iteration $s+1$. The E-steps and M-steps were repeated until the relative difference in the log likelihood from iteration s to iteration $s+1$ was $\leq .00001$.

At each iteration, the log likelihood was computed as

$$\ln(L)^{(s)} = \sum_{i=1}^{N} \ln \sum_{q=1}^{m} \left[\prod_{r=1}^{t} (P_r(\theta_q) | y_{ir}) \right] \pi_q^{(s)}, \tag{5.6}$$

while the relative difference was computed as

$$\left| \frac{\ln(L)^{(s+1)} - \ln(L)^{(s)}}{\ln(L)^{(s)}} \right|. \tag{5.7}$$

Domain scores estimated with this method are labeled EM.

Results

Results from the simulation are summarized in Tables 5.1-5.7. Table 5.1 summarizes the actual observed and estimated domain scores based on simulated item responses to 2,100 students taking ACT's NAEP Science test form. Item responses were simulated for a total of 180 schools: 60 schools with five students taking the test form, 60 schools with 10 students taking the test form, and 60 schools with 20 students taking the test form. Domain scores were estimated for each content area separately, and then for all Science items as a whole. For Physical Science (PS), there were 17 items on ACT's NAEP Science test form; the remaining 45 PS items in the pool were treated as domain items. (The domain excluded the items taken.) Thus, student or group ability was estimated based on the 17 items (items taken), and performance on the remaining 45 items was estimated. For Earth Science (ES), there were 19 items on ACT's NAEP Science test form, with 44 ES domain items. For Life Science (LS), there were 21 items on ACT's NAEP Science test form, with 44 LS domain items. For Science (SCI) overall, there were 57 items on ACT's NAEP Science test form and 133 domain items (190 total items in the Science pool). Domain scores for Science were also computed by weighting the domain score estimate for each content area by the framework weight (0.3*PS + 0.3*ES + 0.4*LS). This weighting is comparable to how plausible values for Science are currently computed.

Table 5.1 Actual and Estimated Domain Scores Based on ACT's NAEP Science Test Form, N = 2,100*

				Method				
Content	# Taken	# Domain	Statistic	EAP (Orig)	EAP (Est)	EM (Orig)	EM (Est)	ACTUAL
PS	17	45	Mean	.40	.39	.41	.40	.41
			90%	.61	.60	.64	.62	.65
			70%	.46	.46	.50	.48	.49
			50%	.38	.37	.40	.38	.40
			30%	.30	.30	.30	.30	.30
			10%	.23	.23	.20	.20	.20
ES	19	44	Mean	.35	.35	.36	.35	.36
			90%	.57	.56	.59	.59	.59
			70%	.42	.41	.42	.42	.43
			50%	.33	.32	.35	.33	.33
			30%	.26	.25	.27	.25	.24
			10%	.18	.18	.16	.16	.15
LS	21	44	Mean	.34	.34	.34	.34	.34
			90%	.49	.49	.51	.51	.52
			70%	.39	.39	.41	.39	.40
			50%	.32	.32	.33	.33	.34
			30%	.37	.26	.27	.27	.27
			10%	.21	.21	.18	.18	.18
SCI	57	133	Mean	.37	.36	.37	.37	.37
			90%	.58	.57	.58	.58	.58
			70%	.44	.43	.45	.43	.44
			50%	.35	.34	.35	.35	.35
			30%	.27	.27	.27	.27	.28
			10%	.19	.19	.18	.18	.18
Frame Work Weight	57	133	Mean	.36	.36	.37	.36	.37
			90%	.55	.54	.57	.57	.58
			70%	.42	.42	.44	.43	.44
			50%	.34	.34	.36	.35	.36
			30%	.32	.27	.28	.27	.27
			10%	.21	.21	.18	.18	.18

*The Domain Excludes Items Taken, and the Domain Score Estimates are Based on Both Original Parameters (Orig) and Estimated Parameters (Est) for the Items Taken.

The domain score estimates from the EAP and EM methods were computed using two sets of parameters: the original parameters used to generate the item responses (labeled *Orig* in Table 5.1), and

calibrated item parameters (for the items taken only[1]) estimated from the item responses (labeled *Est* in Table 5.1). In the latter case, the original parameters were used for the domain items. Multilog 6.1 (Thissen, 1991) was used to calibrate the item parameters for the form taken[2]. The values listed in the column labeled ACTUAL are the actual observed domain scores, computed as the proportion of total points obtained by the group. The average proportion of points for students within the group is summarized in the row labeled 'Mean'. The remaining rows (90%, 70%, 50%, 30%, 10%) give the domain score (proportion of points) for the individual who is at the Nth percentile (i.e., when the 2,100 students were sorted by estimated domain score, 90% of the students scored at or below the individual). The percentiles were computed based on position only; no interpolation methods were used. Note that for the EM method, each ability point on the ability distribution was translated to the proportion of points scale, and the estimated proportion at each ability point was multiplied by the number of total observations to determine the number of observations with the corresponding domain score estimate.

The results for the original and estimated item parameters were very similar for both the mean and the percentiles, indicating that the estimated parameters (reparameterized) for the items taken were very close to the original parameters. Provided the IRT models used fit the data, item calibration (where accurate) should pose little problem to the accuracy of the domain score estimates. If the IRT models are not appropriate, calibration may lead to some loss of accuracy in the domain score estimates. Domain score estimates for PS, ES, LS, and SCI computed by the EM method were very similar to the actual domain scores in terms of both the mean and percentiles. Domain score estimates for PS, ES, and LS

---

[1]An attempt to simultaneously calibrate the items taken and the domain items was unsuccessful because there were too few observations for the number of total items to be calibrated. Operationally the calibration of the domain pool should pose less of a problem because individuals will never take the entire domain.

[2]Multilog uses a different parameterization for the generalized partial credit model than does the special version Bilog-Parscale program currently used for NAEP item calibration. This necessitated reparameterizing the Multilog parameter estimates for polytomous items to be in the form of the original parameters. We used this approach because we did not have access to the special version Bilog-Parscale program and attempts to calibrate the item parameters using Parscale 2.2 (Muraki & Bock, 1991) posed some problems, due to the nature of some items (i.e., very difficult, but not very discriminating). Our work with Parscale suggests that for future operational calibration it may be necessary to write our own item calibration software, use a revised version of Parscale, or use an alternate program such as Multilog.

computed by the EAP method were similar to the actual domain scores for the mean, but showed some slight shrinkage inward at the 90th and 10th percentile points. Students with scores at either extreme were pulled toward the group, so that the estimated proportion correct was higher than the actual proportion correct at the 10th percentile and lower than the actual proportion correct at the 90th percentile. This finding likely results from inward shrinkage of individual EAP estimates of ability. Some shrinkage for the EAP method is also apparent when the Science domain scores are computed by weighting the domain score estimate for each content area by the framework weight. For EAP estimates based on all Science items, the shrinkage is not apparent. This finding suggests that if the EAP method is to be used, it would be better to compute domain score estimates for all Science items as a whole, rather than weighting the estimates for each content area. The EM method appears to be unaffected by inward shrinkage, for the individual content areas or Science estimates computed using framework weights.

Table 5.2 summarizes the domain score estimates based on simulated item responses for 2,100 students taking NAEP Science test form S208D, consisting of blocks F, H, and J. The 2,100 students are the same as reported in Table 5.1. As in Table 5.1, the domain items were treated as mutually exclusive from the items taken (8 PS items taken, 54 PS domain items; 6 ES items taken, 57 ES domain items; 18 LS items taken, 47 LS domain items; and 32 SCI items taken overall, 158 domain items). The domain score estimates reported for this form are based on the original parameters only.

Table 5.2 Actual and Estimated Domain Scores Based on NAEP Test Form (blocks F, H, J), N = 2,100*

Content	# Taken	# Domain	Statistic	Method		
				EAP	EM	ACTUAL
PS	8	54	Mean	.41	.43	.42
			90%	.59	.67	.66
			70%	.45	.52	.51
			50%	.38	.42	.40
			30%	.33	.31	.31
			10%	.31	.20	.21
ES	6	57	Mean	.36	.38	.37
			90%	.47	.61	.61
			70%	.38	.46	.45
			50%	.37	.35	.36
			30%	.31	.27	.26
			10%	.22	.17	.17
LS	18	47	Mean	.36	.37	.37
			90%	.51	.53	.54
			70%	.41	.42	.42
			50%	.35	.35	.35
			30%	.30	.30	.30
			10%	.24	.23	.22
SCI	32	158	Mean	.39	.39	.39
			90%	.59	.60	.60
			70%	.45	.47	.46
			50%	.37	.37	.37
			30%	.30	.29	.30
			10%	.22	.21	.20
Frame Work Weight	32	158	Mean	.38	.39	.39
			90%	.52	.60	.60
			70%	.41	.46	.46
			50%	.37	.37	.37
			30%	.31	.29	.29
			10%	.26	.20	.20

*The Domain Excludes Items Taken, and the Domain Score Estimates are Based on Original Parameters.

The EM-based domain score estimates (mean and percentiles) for each content area, Science as a whole, and Science computed via framework weighting of individual content areas were very similar to the actual domain scores. The EAP-based domain score estimates were also similar to the actual domain scores in terms of the mean score, but showed inward shrinkage at the extreme percentile points.

The inward shrinkage is most apparent for the content areas PS and ES, both of which consist of very small numbers of items taken. The inward shrinkage occurred to a greater degree for the PS and ES content areas on this form than on ACT's NAEP Science test form. Although the EAP method appears to be negatively affected by the number of items taken, the EM method does not appear to be adversely affected by small numbers of items taken.

Table 5.3 summarizes the domain score estimates based on the simulated item responses of the 2,100 students taking ACT's NAEP Science test form. The results differ from results presented in Table 5.1 in that the items taken were treated as a subset of the domain, rather than mutually exclusive from the domain. The domain score estimates in Table 5.3 are based on estimated item parameters for the items taken. There is some expectation that including items taken in the domain may lead to improved domain score estimates if the domain becomes more similar in difficulty to the difficulty of the items taken. In this case, however, the domain score estimates look no closer to the actual domain scores than those reported in Table 5.1.

Table 5.3 Actual and Estimated Domain Scores Based on ACT's NAEP Science Test Form, N = 2,100*

Content	# Taken	# Domain	Statistic	Method		
				EAP	EM	ACTUAL
PS	17	62	Mean	.41	.41	.42
			90%	.60	.62	.64
			70%	.46	.48	.50
			50%	.38	.39	.40
			30%	.32	.31	.32
			10%	.26	.22	.22
ES	19	63	Mean	.37	.37	.38
			90%	.58	.61	.60
			70%	.43	.44	.45
			50%	.34	.35	.35
			30%	.27	.27	.27
			10%	.20	.18	.18
LS	21	65	Mean	.35	.35	.36
			90%	.50	.53	.53
			70%	.40	.40	.42
			50%	.33	.34	.34
			30%	.28	.28	.28
			10%	.23	.20	.19
SCI	57	190	Mean	.38	.38	.38
			90%	.58	.59	.59
			70%	.44	.44	.45
			50%	.36	.36	.37
			30%	.29	.29	.29
			10%	.21	.20	.21
Frame Work Weight	57	190	Mean	.37	.37	.38
			90%	.55	.58	.58
			70%	.43	.44	.45
			50%	.35	.36	.36
			30%	.29	.29	.29
			10%	.23	.20	.20

*The Domain Includes Items Taken, and the Domain Score Estimates are Based on Estimated Parameters for the Items Taken.

Table 5.4 summarizes domain score estimates for ACT's NAEP Science test form based on a subset of schools from the 180 total schools. The number of observations within this grouping of schools is 200. For the PS, ES, and LS content areas, some inward shrinkage is again apparent for the EAP

method at the extreme percentiles. The EAP method appears to be affected by both small numbers of items taken and smaller numbers of observations per group (although the average estimated domain scores for the group are similar to the actual observed average). The EM method does not appear to display similar shrinkage. Thus, it is likely that the EM method could be realistically applied to smaller groupings, such as a school district, if domain score estimates were desired at that level. However, when reporting results for groups as small as 200, it may be best to report means only, because the percentiles may be less accurate to some degree.

Table 5.4 Actual and Estimated Domain Scores Based on ACT's NAEP Science Test Form, N = 200 (Schools 6-45)*

Content	# Taken	# Domain	Statistic	Method		
				EAP	EM	ACTUAL
PS	17	45	Mean	.41	.43	.43
			90%	.53	.66	.64
			70%	.48	.52	.51
			50%	.40	.44	.42
			30%	.32	.35	.33
			10%	.23	.20	.21
ES	19	44	Mean	.36	.37	.37
			90%	.58	.55	.59
			70%	.43	.42	.46
			50%	.35	.37	.37
			30%	.28	.30	.27
			10%	.19	.15	.15
LS	21	44	Mean	.35	.36	.35
			90%	.49	.50	.53
			70%	.41	.44	.40
			50%	.33	.35	.34
			30%	.28	.28	.29
			10%	.22	.21	.18
SCI	57	133	Mean	.38	.39	.39
			90%	.59	.56	.59
			70%	.45	.45	.45
			50%	.37	.36	.37
			30%	.30	.29	.30
			10%	.20	.19	.19
Frame Work Weight	57	133	Mean	.37	.38	.38
			90%	.53	.56	.58
			70%	.44	.46	.45
			50%	.36	.38	.37
			30%	.29	.31	.30
			10%	.21	.19	.18

*The Domain Excludes Items Taken, and the Domain Score Estimates are Based on Estimated Parameters for the Items Taken.

Table 5.5 summarizes domain score estimates based on simulated item responses to 7,875 students taking ACT's NAEP Science test form. Item responses were simulated for a total of 135 schools: 45 schools with 25 students taking the test form, 45 schools with 50 students taking the test form, and 45 schools with 100 students taking the test form. The number of total observations is similar to the final

number typically summarized for the national NAEP. Operationally, all individuals in the national sample would probably not take the same form. Because of the IRT invariance principle, however, both the EM and EAP methods could be applied to the group as a whole, regardless of which form is taken, provided the item parameters are on the same scale across forms taken, and the domain is the same for each form taken. If each form meets the NAEP framework (i.e., is like ACT's NAEP Science test form), we can expect national results to be similar in terms of accuracy of the domain score estimates, regardless of whether everyone in the nation takes a single form or multiple forms. Comparing the results to Table 5.1 suggests that the estimation methods are not necessarily advantaged by having close to 8,000 observations versus 2,100 observations.

Table 5.5 Actual and Estimated Domain Scores Based on ACT's NAEP Science Test Form, N = 7,875*

Content	# Taken	# Domain	Statistic	Method		
				EAP	EM	ACTUAL
PS	17	45	Mean	.39	.40	.41
			90%	.60	.62	.65
			70%	.45	.48	.50
			50%	.37	.38	.38
			30%	.30	.30	.30
			10%	.24	.21	.20
ES	19	44	Mean	.35	.35	.36
			90%	.56	.59	.59
			70%	.41	.42	.43
			50%	.32	.33	.34
			30%	.25	.25	.25
			10%	.18	.17	.16
LS	21	44	Mean	.34	.34	.34
			90%	.49	.51	.52
			70%	.38	.39	.40
			50%	.32	.33	.32
			30%	.27	.27	.26
			10%	.21	.19	.18
SCI	57	133	Mean	.36	.37	.37
			90%	.56	.56	.59
			70%	.43	.43	.44
			50%	.34	.35	.35
			30%	.27	.27	.27
			10%	.20	.19	.19
Frame Work Weight	57	133	Mean	.36	.36	.37
			90%	.54	.57	.58
			70%	.41	.43	.44
			50%	.34	.35	.34
			30%	.27	.27	.27
			10%	.21	.19	.18

*The Domain Excludes Items Taken, and the Domain Score Estimates are Based on Estimated Parameters for the Items Taken.

Setting Standards for NAEP Via Domain Scores

If domain scores are adopted for NAEP, standards could be determined for domain performance as the proportion of possible domain points that Basic, Proficient, and Advanced students should receive if they were to take the domain. Although standard-setting procedures may not change, the process of setting standards may be an easier task for the standard setters when standards are based on a clearly defined pool of items. ACT is planning to evaluate standard-setting techniques where experts decide what score a borderline Advanced, Proficient, or Basic student would obtain for each item in the pool. From this information, an expected domain score could be computed in terms of the proportion of total points the borderline student would obtain. [Note that this, and other weighting options, were discussed in Chapter Three.] Standards, then, would be set in terms of the proportion of total possible domain points a student classified as Advanced, Proficient, or Basic should obtain. Such a process could be very tedious if the domain item pool were large. However, standards could be determined from a subset of domain items, if the subset were representative of the domain at each level for which domain scores were to be reported (i.e., so that someone who receives X proportion of points on the PS, ES, LS, or SCI items within the subset could expect to get X proportion of points on the domain also).

Table 5.6 summarizes the domain score estimates of the simulation in terms of the percentage of individuals who would be classified as Advanced, Proficient, or Basic for some arbitrary cutscores. Arbitrary cutscores were used, because NAGB has not yet decided upon final cutscores for the 1996 NAEP Science Assessment. The table is based on the 2,100 students taking ACT's NAEP Science test form, where estimated parameters were used for the items taken, and the domain excluded the items taken. The cutscores for each content area were chosen to be .65 proportion of total points for Advanced, .50 proportion of total points for Proficient, and .35 proportion of total points for Basic. In practice, the cutscores may differ for each content area according to the various difficulties of the content areas.

For the cutscores employed, Table 5.6 shows that both the EM and the EAP method consistently underestimated the number of individuals classified in each achievement level. Typically, however, the

number classified by the EM method more closely approached the actual observed number than did the number classified by the EAP method. The underestimation for the EAP method likely results from inward shrinkage of individual EAP estimates of ability. The underestimation for the EM method is probably due to the fact that only a fixed number of score points is possible, one for each quadrature point employed in the estimation.

Table 5.6 N-Counts and Percents ≥ Achievement Levels Based on ACT's NAEP Science Test Form, N = 2,100*

		Achievement Level					
		Basic		Proficient		Advanced	
Content	Method	N	Percent	N	Percent	N	Percent
PS	EAP	1173	.56	466	.22	110	.05
	EM	1192	.57	618	.29	164	.08
	ACTUAL	1204	.57	619	.29	212	.10
ES	EAP	895	.43	327	.16	69	.03
	EM	920	.44	385	.18	98	.05
	ACTUAL	983	.47	418	.20	121	.06
LS	EAP	831	.40	178	.08	16	.01
	EM	932	.44	216	.10	29	.01
	ACTUAL	957	.46	271	.13	39	.02
SCI	EAP	1009	.48	385	.18	78	.04
	EM	1014	.48	415	.20	89	.04
	ACTUAL	1068	.51	430	.20	99	.05

*The Domain Excludes Items Taken, and the Domain Score Estimates are Based on Estimated Parameters for the Items Taken. Advanced = .65 proportion of points, Proficient = .50 proportion of points, and Basic = .35 proportion of points.

Table 5.7 summarizes the percentage of individuals who would be classified as Advanced, Proficient, or Basic based on the 7,875 observations summarized in Table 5.5. The results are similar to the results based on the 2,100 observations. Even for the larger sample size, the underestimation of numbers classified into each achievement level is apparent for both methods. When the domain score estimates are evaluated in terms of percentages at each achievement level (Tables 5.6-5.7), the estimated

percentages seem to vary from the actual percentages somewhat more than when the domain score estimates are evaluated in terms of the estimated domain score at a percentile point (Tables 5.1-5.5). For the EAP method, the inward shrinkage of the domain score estimates is even more apparent when the domain scores are evaluated according to percentages at each achievement level. In addition, some shrinkage is also apparent with the EM method in terms of achievement level classification. The underestimation displayed for the EAP and EM methods may vary somewhat for different cutscores. Also, it may be possible to incorporate student and school background variables (i.e., measures of parent education, school socioeconomic status) into the analyses to improve the precision of the domain score estimates. One possible procedure is discussed in the section, *Incorporating Background Variables Into Analyses*, that follows.

Table 5.7 N-Counts and Percents ≥ Achievement Levels Based on ACT's NAEP Science Test Form, N = 7,875*

		Achievement Level					
		Basic		Proficient		Advanced	
Content	Method	N	Percent	N	Percent	N	Percent
PS	EAP	4341	.55	1741	.22	458	.06
	EM	4385	.56	2199	.28	647	.08
	ACTUAL	4509	.57	2386	.30	813	.10
ES	EAP	3332	.42	1260	.16	310	.04
	EM	3405	.43	1468	.19	396	.05
	ACTUAL	3704	.47	1599	.20	513	.07
LS	EAP	3041	.39	682	.09	88	.01
	EM	3464	.44	791	.10	147	.02
	ACTUAL	3604	.46	995	.13	173	.02
SCI	EAP	3730	.47	1385	.18	307	.04
	EM	3714	.47	1497	.19	370	.05
	ACTUAL	4001	.51	1613	.20	417	.05

*The Domain Excludes Items Taken, and the Domain Score Estimates are Based on Estimated Parameters for the Items Taken. Advanced = .65 proportion of points, Proficient = .50 proportion of points, and Basic = .35 proportion of points.

Measuring Trends Over Time Via Domain Scores

NAEP currently uses a separate assessment program to monitor trends; this assessment is based on curricular specifications of the 1970s, and has remained unchanged since the 1970s. A domain score approach to NAEP would facilitate the tracking of trends over time by possibly eliminating the need for the two separate testing programs that are currently in use. Domain scores as defined in this report are simple percentages, representing the percentage of possible domain points the group is expected to achieve. Thus, if the domain remains fixed over a period of years, the definition of proficiency remains constant and performance can be compared across years. If assessment specifications and administration procedures remain constant for a sufficient period of time, content on a new form would have the same meaning as in the content domain, and domain scores would also have the same meaning over time. This method seems especially promising for NAEP given plans to maintain NAEP content and a fixed domain for at least 10 years (Forsyth, Hambleton, Linn, Mislevy, & Yen, 1996).

ACT's proposed design would keep test specifications and administration procedures constant over different test administrations. Either common item equating procedures or random groups equating could be used to link the forms over time to the domain scores that were initially established.

When changes in test specifications are necessitated by curricular changes, then the content domain will also change. If the change in the domain is only in difficulty, for the purpose of tracking trends, an adjustment could be made to the reported domain scores, to reflect the change in domain difficulty. The linkage could be made using either common items between the two assessments or by randomly assigning examinees to take the new and old assessments. These methods assume that the new and old assessments measure the same constructs, and differ only in difficulty.

When the domains change in meaning and forms across domains are no longer measuring the same construct, then some concordance procedure might be used to link the assessments by randomly assigning examinees to take the new and old assessments. The concordance would need to be verified in various subgroups, as concordance relationships typically are subgroup specific.

Incorporating Background Variables Into Analyses

The results from the simulation suggest that for some conditions, such as small numbers of items taken, it may be desirable to improve the precision of the domain scores estimated by the EAP and EM methods. It may be possible to improve the precision of the domain score estimates by incorporating background variables into analyses. Background variables are currently employed in NAEP analyses to get estimates of ability distributions for students, from which plausible values are drawn. A simpler method of incorporating background variables was investigated via a simulation study, with the goal of easing the complexity and computational burden of methods currently employed in the computation of plausible values and alternate methods, such as those discussed in Adams, Wilson, and Wu (1997).

Overview

Using real data from a special achievement level setting validation study (ACT, 1995), a prediction model was chosen, where student's observed proportion correct scores were predicted. The variables in the model were chosen to maintain model parsimony with a suitable magnitude of prediction. Data were then simulated to represent the real data and the model, in terms of the relationship between variables and the mean square error of the model. A true latent ability was computed for each examinee, as the sum of a predicted score from the model and an error term. Item responses were generated using this latent ability. The prediction model was then fit to the simulated data to get a predicted score for each examinee. The predicted scores were linked to both the EAP estimates and the EM ability distribution to create new domain score estimates combining the IRT-based estimates with the estimates from the prediction model.

Method

The real student-level data consisted of item responses to four NAEP Geography blocks and responses to the student questionnaire. Nineteen separate classrooms were included in the study; for each participating classroom, responses to the teacher questionnaire and the school characteristics and policies questionnaire were also available. Classroom sizes ranged from 11 to 64, for a total of 670 students. Because the classrooms were selected as part of a special study, it is possible that the students represented in the study were more motivated than the typical NAEP-tested student.

In fitting a prediction model, only the student-level and school variables were considered for inclusion in the model. The form of many of the items in the teacher questionnaire did not appear suitable for modeling. In addition, because data were to be simulated for the background variables in the prediction model, dichotomous and categorical variables were not considered for inclusion in the model. Thus, only ordered categorical variables with three or more categories were evaluated. Because the student-level data was nested within schools, multilevel models were employed to incorporate the hierarchical structure of the data.

Multilevel models incorporate variables measured at different levels of analysis, employing a submodel for each level of measurement. The multilevel model corresponding to the NAEP structure is a two-level model, where students are represented at Level 1 (the within-school model) and the school districts are represented at Level 2 (the between-school model). The Level-1 model measures variation within schools:

$$Y_{is} = \beta_{0s} + \beta_{1s}X_{1is} + \beta_{2s}X_{2is} + \ldots + \beta_{vs}X_{vis} + \varepsilon_{is} \qquad (5.8)$$

for $s = 1,2, \ldots, S$ schools. X_{vis} is a measured characteristic of student i within school s, β_{vs} represents the expected change in the outcome for a fixed unit of X_v for student i within school s, and ε_{is} is random error.

In the multilevel model, each parameter represented in the Level 1 model is allowed to vary across the units of analysis at Level 2. This variation is represented in an equation modeling the intercepts and slopes as outcomes at the second level. The Level 2 model is a between-schools model:

$$\beta_{vs} = \gamma_{v0} + \gamma_{v1}Z_{1s} + \gamma_{v2}Z_{2s} + \ldots + \gamma_{vw}Z_{ws} + u_{vs} \qquad (5.9)$$

where Z_{ws} is a measured characteristic of the school, γ_{vw} represents the effect of Z_{ws} on the *vth* parameter for school *s*, and u_{vs} is random error. More detail about the structure of multilevel models is available in Bryk and Raudenbush (1992).

The multilevel models were fit using the program HLM 2.2 (Bryk, Raudenbush, Seltzer, & Congdon, 1989). HLM computes empirical Bayes (EB) estimates of the first level parameters. The estimates are said to be optimal in that no other estimates have a smaller expected mean-squared error. The estimates are often called shrinkage estimates because the ordinary least squares regression line for each school is pulled toward a predicted value based on the school-level model. When no second level predictors are included in the model, shrinkage is toward a common regression line. The amount of shrinkage is conditional on the reliability of the first level parameters; when the first level parameters are reliable, shrinkage is minimal.

The final model employed in the study contained two student-level predictors and one school-level predictor:

$$\begin{aligned} \text{PROP} &= \beta_0 + \beta_1(\text{PERFORM}) + \beta_2(\text{EDUC}) + \varepsilon \\ \beta_0 &= \gamma_{00} + \gamma_{01}(\text{LUNCH}) + u \\ \beta_1 &= \gamma_{10} + \gamma_{11}(\text{LUNCH}) + u \\ \beta_2 &= \gamma_{20} + \gamma_{21}(\text{LUNCH}) + u \end{aligned} \qquad (5.10)$$

The student-level variables consisted of a motivation question addressing the student's perception of their performance on the test (PERFORM), and the self-reported average education level of the student's mother and father (EDUC). The school-level variable was the percentage of students within the school receiving a subsidized school lunch and/or nutrition program (LUNCH). The outcome variable in the model was

the observed proportion correct on the four NAEP Geography blocks (PROP). The model was chosen because it was parsimonious, yet accounted for a suitable amount of variance in the proportion correct score. A regression model of the first level variables fit to the entire group accounted for 35% of the variance in the proportion correct score. Because the students in this special study may have been more motivated than the typical NAEP-tested student, the performance of the variables in predicting proportion correct scores may not have been typical. Also, the students responded to four out of seven possible Geography blocks, which likely provided a very good estimate of their domain performance.

After fitting the HLM model to the real data, data were simulated to represent the real data and model as closely as possible. Item responses and background variables were generated for students within nineteen schools. Each school was assigned a sample size and value of the school-level variable (LUNCH) as observed in the real data. For each student within a school, values for the two student-level variables (PERFORM and EDUC) were drawn at random from a N(0,1) distribution. The generated variables were rescaled to match the means and covariances observed among the real variables, using standard techniques. Note that although the student-level background variables were discrete ordered categorical variables in the real data, the generated values were not discretized in the simulation.

Using the parameter estimates from the HLM model in the real data, and the generated background variables, a predicted true score was computed for each examinee. An error term was also generated for each examinee by drawing a random value from a N(0,MSE) distribution, where MSE was the mean square error from the multilevel model fit to the real data. A true domain score was computed for each examinee as the sum of the predicted true score and the error term. The error term was added to give the domain score the variability that would be expected in real data. The true domain scores were standardized and item responses were generated for each examinee from the standardized latent ability measure and the item parameters in the real data. Item responses were generated for two separate Geography test forms. One form was an actual NAEP Geography form (blocks C and H). The second form was created to match the NAEP Geography framework in terms of percentage of testing time allotted

to each content area, by item type. Note that this form was not intended to match the Geography framework in terms of any dimensions other than content specifications. The Geography item pool did not undergo any of the rigorous test development processes discussed in Chapter Two.

After the simulated data was complete, the HLM model was refit to the simulated data to obtain parameter estimates for each school. The outcome variable in the model was the unstandardized true domain score of the simulation. A predicted score was computed for each student within each school based on their background variables and the parameter estimates from the multilevel model. The predicted scores were linked to both the EAP estimates and the EM ability distribution to create new domain score estimates combining the IRT-based estimates with the estimates from the prediction model. For each student, the predicted score was linked to their EAP domain score estimate by computing a weighted average of the two estimates, where each estimate was weighted by the inverse of its variance. In linking the predicted scores to the EM ability distribution, the predicted proportion correct scores were mapped to the theta scale of the ability distribution, so that each distribution was on the same scale. At each theta point, a weighted average of the N-count for the EM distribution and the predicted score distribution was computed, where each N-count was weighted by the inverse of the variance of the respective distribution. Thus, a weighted ability distribution was computed on the theta scale. Each ability point in this distribution was then translated back to the proportion of points scale, and the estimated proportion at each ability point was multiplied by the number of total observations to determine the number of observations with the corresponding domain score estimate.

Results

The results of the simulation are summarized in Tables 5.8-5.11. All domain score estimates in the tables are based on the original item parameters used to generate the item responses. Table 5.8 gives the actual and estimated domain scores based on simulated item responses to the actual NAEP Geography form consisting of blocks C and H. Domain scores were estimated for each content area separately, and for all Geography items as a whole. The items on the form taken were not included in the domain. The

results are summarized for the content areas Space and Place (SP; 17 items taken, 41 domain items), Environment and Society (ES; 6 items taken, 26 domain items), and Spatial Dynamics and Connections (SDC; 11 items taken, 23 domain items). Overall there were 34 Geography (GEO) items taken and 90 separate Geography domain items. The columns labeled EAP, EM, and ACTUAL contain domain scores computed as discussed in the previous section, *Projecting to a Domain*. The columns labeled EAP2 and EM2 contain the results from the combined IRT estimates and the predicted estimates from the multilevel model[3]. The column labeled TRUE contains the domain score estimates that result when true latent ability (the standardized latent ability used to generate the item responses) is applied in Equation 5.2.

[3]The results from the prediction model reported in these tables are actually based on ordinary least squares (OLS) regression coefficients from a regression model fitted to each school separately, rather than the empirical Bayes (EB) estimates discussed earlier. Although one advantage of EB estimates is reduced variance, in this case, the shrinkage due to the EB estimates was so great that when the predicted scores were combined with the EAP domain scores, even greater shrinkage occurred than observed with the EAP scores alone. The student-level variables showed very low reliability, so that a substantial amount of shrinkage occurred in the EB estimates. The low reliability was probably due to the small number of response categories possible in the variables. Because one goal was to increase the variability of the predicted domain scores, the OLS estimates were considered more suitable in this application.

Table 5.8 Actual and Estimated Domain Scores Based on the NAEP Geography Form (Blocks C and H)*

Content	# Taken	# Domain	Statistic	Method					
				EAP	EAP2	EM	EM2	TRUE	ACTUAL
SP	17	41	Mean	.55	.52	.55	.54	.55	.55
			90%	.80	.78	.81	.83	.80	.81
			70%	.67	.65	.68	.68	.68	.68
			50%	.55	.54	.54	.54	.56	.55
			30%	.44	.40	.43	.43	.43	.43
			10%	.30	.24	.28	.26	.29	.28
ES	6	26	Mean	.36	.37	.38	.39	.38	.38
			90%	.60	.63	.64	.67	.64	.66
			70%	.40	.44	.44	.47	.46	.46
			50%	.32	.34	.35	.35	.34	.34
			30%	.25	.26	.26	.26	.25	.24
			10%	.21	.18	.19	.18	.17	.15
SDC	11	23	Mean	.39	.40	.40	.41	.40	.40
			90%	.66	.67	.65	.71	.65	.69
			70%	.47	.49	.49	.49	.49	.50
			50%	.36	.38	.38	.38	.37	.36
			30%	.27	.27	.28	.28	.27	.25
			10%	.20	.16	.17	.16	.17	.17
GEO	34	90	Mean	.45	.44	.45	.45	.45	.45
			90%	.71	.71	.70	.73	.70	.72
			70%	.54	.55	.56	.56	.55	.56
			50%	.42	.44	.42	.44	.43	.43
			30%	.33	.32	.34	.32	.33	.32
			10%	.22	.18	.21	.20	.21	.20

*The Domain Excludes Items Taken, and the Domain Score Estimates are Based on the Original Parameters for the Items Taken.

Comparing the results for true latent ability (TRUE) with the actual observed domain score results (ACTUAL) shows that there is some shrinkage inherent in using an IRT model to compute the proportion correct for the group or for each student. The proportion correct at the 90th and 10th percentiles are typically less extreme for TRUE than for ACTUAL. The results for the EM method follow the results for TRUE pretty closely. The EAP method typically shows greater shrinkage than TRUE, particularly for the ES content area (6 items taken). The results for EAP2 and EM2 both show a stretching of the

distribution at the tails. Typically, more students are assigned extreme scores under these methods (see proportion correct at 90% and 10%) than under the EAP or EM methods alone. However, in many cases, the estimated proportions correct at the 90th and 10th percentiles are more extreme than those for the actual observed domain scores and the stretching sometimes occurs only in one tail.

Table 5.9 summarizes the domain score estimates of the simulation in terms of the percentage of individuals who would be classified as Advanced, Proficient, or Basic for arbitrary cutscores of .75, .60, or .45 proportion of total points, respectively. The results for TRUE again show a tendency for the IRT-based methods to slightly underestimate the actual number classified in each achievement level. The results for EM are very similar to the results for TRUE, in many cases. The results for EAP show even more underestimation of ACTUAL than for TRUE. The EAP2 and EM2 methods show some tendencies to raise the N-counts in each achievement level closer to the ACTUAL N-counts. In some cases, however, they raise the N-counts beyond those in ACTUAL.

Table 5.9 N-Counts and Percents ≥ Achievement Levels Based on the NAEP Geography Form (Blocks C and H)*

		Achievement Level					
		Basic		Proficient		Advanced	
Content	Method	N	Percent	N	Percent	N	Percent
SP	EAP	447	.67	289	.43	114	.17
	EAP2	424	.63	265	.40	89	.13
	EM	437	.65	280	.42	118	.18
	EM2	425	.63	273	.41	117	.17
	TRUE	455	.68	289	.43	116	.17
	ACTUAL	437	.65	280	.42	124	.19
ES	EAP	156	.23	73	.11	26	.04
	EAP2	189	.28	83	.12	22	.03
	EM	186	.28	84	.13	25	.04
	EM2	201	.30	98	.15	32	.05
	TRUE	216	.32	90	.13	24	.04
	ACTUAL	222	.33	109	.16	33	.05
SDC	EAP	222	.33	87	.13	37	.06
	EAP2	251	.37	96	.14	33	.05
	EM	237	.35	90	.13	41	.06
	EM2	254	.38	108	.16	50	.07
	TRUE	240	.36	96	.14	37	.06
	ACTUAL	242	.36	114	.17	47	.07
GEO	EAP	302	.45	149	.22	45	.07
	EAP2	322	.48	154	.23	43	.06
	EM	305	.46	159	.24	38	.06
	EM2	312	.47	170	.25	47	.07
	TRUE	308	.46	150	.22	45	.07
	ACTUAL	321	.48	156	.23	52	.08

*The Domain Excludes Items Taken, and the Domain Score Estimates are Based on Original Parameters for the Items Taken. Advanced = .75 proportion of points, Proficient = .60 proportion of points, and Basic = .45 proportion of points.

The results for the form created to match the NAEP Geography frameworks are given in Tables 5.10-5.11. Table 5.10 summarizes the actual and estimated domain scores, while Table 5.11 summarizes the N-counts and percents falling into each achievement level classification. Results are similar for this form, as for the actual NAEP Geography form (blocks C and H), in that results based on the IRT model show some shrinkage of extreme scores toward the group (TRUE vs. ACTUAL). Results for the EM method pretty closely match results based on the true latent ability, while the EAP method shows greater shrinkage than TRUE or EM. The EAP2 and EM2 methods both show a reduction in the shrinkage that occurs at the 90th and 10th percentiles. Again, they result in more students receiving extreme scores, but tend to overcompensate at the tails. In addition, the methods may spread more students into one tail, but not the other.

Table 5.10 Actual and Estimated Domain Scores Based on the Created Geography Form*

Content	# Taken	# Domain	Statistic	Method					
				EAP	EAP2	EM	EM2	TRUE	ACTUAL
SP	15	43	Mean	.59	.55	.59	.57	.59	.59
			90%	.82	.80	.83	.83	.83	.85
			70%	.71	.68	.72	.72	.72	.73
			50%	.59	.55	.61	.58	.60	.58
			30%	.47	.43	.46	.44	.47	.46
			10%	.33	.26	.32	.27	.31	.29
ES	11	21	Mean	.39	.39	.40	.40	.40	.40
			90%	.64	.66	.66	.71	.66	.69
			70%	.46	.48	.49	.49	.49	.49
			50%	.34	.37	.37	.40	.37	.37
			30%	.26	.27	.27	.27	.28	.26
			10%	.21	.18	.19	.17	.18	.17
SDC	11	23	Mean	.42	.42	.43	.44	.43	.44
			90%	.71	.71	.73	.76	.70	.71
			70%	.50	.51	.54	.54	.54	.54
			50%	.37	.40	.40	.42	.41	.40
			30%	.29	.28	.30	.28	.30	.31
			10%	.23	.19	.19	.18	.19	.17
GEO	37	87	Mean	.48	.47	.48	.48	.48	.49
			90%	.75	.74	.76	.76	.74	.75
			70%	.59	.59	.60	.60	.60	.60
			50%	.46	.47	.48	.48	.47	.47
			30%	.36	.34	.36	.36	.36	.36
			10%	.24	.19	.23	.21	.23	.23

*The Domain Excludes Items Taken, and the Domain Score Estimates are Based on the Original Parameters for the Items Taken

Conclusions

The domain scores based on combining the IRT-based estimates with the estimates from the prediction model show some tendencies to compensate for the shrinkage of extreme scores inherent with the IRT approach. The EAP2 and EM2 estimates do tend to result in some students receiving more extreme scores than they would under the IRT model alone, but they do so inconsistently. The overall precision of these new estimates appears to be much less than the precision of the EM estimates relative to the TRUE domain scores. Overall, the EAP2 results appear no better than the EM results, and although

the EM2 method does compensate somewhat for the shrinkage inherent in the IRT approach, it is inconsistent and sometimes compensates to a greater degree than desired.

The EM domain score estimation method alone is somewhat disadvantaged in that once the EM step of the domain score estimation process is completed (i.e., the group ability distribution is estimated), all individual information is lost. Any mapping of the EM ability distribution to individual information is crude at best. Although the EM method without background information displays some shrinkage, it appears to be only to a degree that can be expected by employing an IRT model. This shrinkage is undesirable only to the degree that extreme scores are believed to be representative of an actual extreme ability. If the shrinkage that occurs with the EM method is to be controlled, it may have to be done at the application of the EM algorithm (see Equation 5.4) by computing the proportion correct given item responses and background information. The plausibility of this approach should probably be investigated.

Table 5.11 N-Counts and Percents ≥ Achievement Levels Based on the Created Geography Form*

		Achievement Level					
		Basic		Proficient		Advanced	
Content	Method	N	Percent	N	Percent	N	Percent
SP	EAP	493	.74	330	.49	147	.22
	EAP2	457	.68	290	.43	118	.18
	EM	482	.72	335	.50	159	.24
	EM2	461	.69	318	.47	149	.22
	TRUE	481	.72	333	.50	168	.25
	ACTUAL	486	.73	330	.49	189	.28
ES	EAP	214	.32	84	.13	24	.04
	EAPs	248	.37	98	.15	27	.04
	EM	241	.36	93	.14	25	.04
	EM2	258	.39	113	.17	35	.05
	TRUE	236	.35	103	.15	25	.04
	ACTUAL	253	.38	114	.17	29	.04
SDC	EAP	253	.38	125	.19	45	.07
	EAP2	274	.41	134	.20	48	.07
	EM	279	.42	143	.21	58	.09
	EM2	288	.43	153	.23	65	.10
	TRUE	298	.44	146	.22	50	.07
	ACTUAL	296	.44	152	.23	54	.08
GEO	EAP	349	.52	191	.29	66	.10
	EAP2	358	.53	190	.28	64	.10
	EM	363	.54	207	.31	66	.10
	EM2	363	.54	212	.32	72	.11
	TRUE	362	.54	200	.30	63	.09
	ACTUAL	352	.53	205	.31	69	.10

The Domain Excludes Items Taken, and the Domain Score Estimates are Based on Original Parameters for the Items Taken. Advanced = .75 proportion of points, Proficient = .60 proportion of points, and Basic = .45 proportion of points.

CHAPTER SIX

SUMMARY AND CONCLUSIONS

The core concept of ACT's recommended redesign of NAEP is that an appropriate assessment design can reduce dependence on complex analysis procedures, with the outcome being more stable, useful, and timely assessment results. In our redesign, some of the burden for producing valid results is shifted from the analysis procedures to the assessment design. In addition, we make suggestions about sampling and provide additional psychometric analyses that can enhance the usefulness of NAEP scores. The areas that we studied are summarized in the following paragraphs.

Assessment Design

We considered two alternative assessment designs that depend on the content area to be assessed. In one design, individual forms of the assessment are developed that cover the entire range of content and can be completely administered to examinees. In the second design, individual forms of the assessment are developed that can be completely administered to individual examinees such that the entire range of content can be covered by a group of individual forms. Such a group of individual forms is called a super form.

We used the NAEP Science Framework and NAEP Science items from the 1996 NAEP to construct individual forms consistent with the first of these design alternatives. We found that an individual form could represent the Science Framework. We also found that the Science assessments constructed under the current BIB design do not closely reflect the NAEP Science Framework. In fact, we were able to develop only one individual form from the entire set of test questions that comprise the 1996 Science booklets. To address these shortcomings, we recommend that precise content and statistical specifications for forms be developed and that enhanced pretesting procedures be used.

Scoring

In the current NAEP, the weighting of multiple-choice and constructed response items is based largely on psychometric considerations. We believe that NAEP results would be more readily interpretable and that more meaningful NAEP standards could be set if the weighting of different formats were based much more heavily on content considerations. In our approach, weights for different item types would be clearly specified in the frameworks by policymakers and/or content matter specialists. Psychometric procedures would be used primarily to study the psychometric properties of the weighted scores. A process for establishing weights is described in Chapter Three, and empirical results that support our conclusions are presented.

Scaling and Equating

We suggest using much simpler procedures for scaling and equating. When an individual form can adequately reflect the content framework, simple and straightforward scaling and equating procedures can be used. More complicated scaling and equating procedures would be needed for super forms. Current NAEP procedures assume that item functioning is unaffected by the position of the block in the test booklet. Current NAEP procedures also assume that the particular mix of item types (constructed response versus multiple-choice) that an examinee receives has no effect on the psychometric dimensions measured. ACT's redesigned approach does not rely on these strong, and what we believe to be untenable, assumptions. Instead, we propose keeping the item position within booklet constant and, to the extent possible, administering the same mix of item types to all examinees. We also recommend using psychometric models at the test level, rather than at the item level. For these reasons, our design and analysis procedures can be expected to produce more stable results.

Estimating Score Distributions

When an individual form can adequately reflect the content framework, observed score distributions are suggested for use in reporting results. These observed score distributions could be reported very rapidly using extremely simple methodology. Issues surrounding estimating true score distributions and a relevant simulation study are described in detail in Chapter Three.

Sampling

ACT suggests simplifying the NAEP sampling procedures. We propose that the primary sampling unit be the school rather than a geographic unit, as is presently done. We propose that schools be selected with equal probability, rather than proportional to size, and that all students within schools be tested. We believe that these simplifications make more efficient use of resources and can achieve the same precision while reducing the cost. A simulation study was conducted to study these design issues. We also described a process for combining samples from the State and National Assessments.

Further Psychometric Analyses

We conducted a study of the feasibility of reporting domain scores for NAEP under the proposed design. Methodology developed at ACT for estimating domain scores from the individual form taken was adapted to NAEP conditions, and the suitability of the method for NAEP was examined. We concluded that this domain score methodology has the potential to be used with NAEP to enhance score reporting and that it could be used to enhance the NAEP standard setting process.

We also described how the domain score approach could be used to measure trends, and we considered approaches for adjusting domain scores to reflect changes in domain content as the assessments change. We discussed the use of multilevel models to improve the precision of item-level or domain score-level proportion-correct and the precision of estimated score distributions. We

considered whether the background information currently collected is suitable for multilevel analyses and whether the amount of background information collected can be minimized. We also considered the type of sampling necessary for implementing multilevel models.

Independent Replication

ACT firmly believes that all important NAEP results should be independently replicated. Presently, NAEP scaling and distribution estimation procedures incorporate heuristics that are not described in sufficient detail for independent replication. ACT recommends that this situation be remedied by using much simpler psychometric and statistical methods and by making public all algorithms that are used in the scaling and distribution estimation process.

Conclusions

We have presented an alternative design for NAEP that has the potential to simplify greatly the data analysis procedures needed to produce assessment results. We gain simplicity and stability by relying more on test design and less on statistical assumptions. Our approach requires more careful consideration of test content and constrains the test development process to a greater extent than the current design. However, we believe that our design leads to simpler statistical and psychometric analyses and more stable and useful results. In addition, our design has the potential to provide more timely results using design and analysis procedures that can be readily interpreted and replicated by other psychometricians and testing agencies.

REFERENCES

ACT (1995). *Research studies on the achievement levels set for the 1994 NAEP in Geography and U.S. History*. Unpublished manuscript.

Adams, R.J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*, 47-76.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education, 5*, 225-264.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models*. London: Sage.

Bryk, A.S., Raudenbush, S.W., Seltzer, M., & Congdon, R.T. (1989). *An introduction to HLM: Computer program and user's guide* (2nd ed.). Chicago: University of Chicago Department of Education.

Cochran, W. G. (1977) *Sampling techniques, 3rd edition.* New York: John Wiley & Sons.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B, 39*, 1-38.

Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., & Yen, W. (1996). *Design Feasibility Team Report to the National Assessment Governing Board*. Washington, D.C.: The National Assessment Governing Board.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer-Nijhoff Publishing.

Hanson, B. A. (1991a). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement, 15*, 391-408.

Hanson, B. A. (1991b). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes*. ACT Research Report 91-5. Iowa City, IA: American College Testing.

Hodges, J. S. (1987). Uncertainty, policy analysis, and statistics (with discussion). *Statistical Science, 2*, 259-291.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Research Rep. No. 87-31). Princeton NJ: Educational Testing Service.

Kish, L. (1965) *Survey sampling*. New York: John Wiley & Sons.

REFERENCES, CONT.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28*, 257-282.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Leary, L.F., & Dorans, N.J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*, 387-413.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*, 23-39.

Lord, F. M. (1965). A strong true score theory, with applications. *Psychometrika, 30*, 239-270.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing (An empirical Bayes estimation problem). *Psychometrika, 34*, 259-299.

Mollenkopf, W.G. (1950). An experimental study of the effects on item analysis data of changing item placement and test-time limit. *Psychometrika, 15*, 291-315.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E., & Bock, R.D. (1991). *Parscale: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks*. Chicago, IL: Scientific Software, Inc.

NAGB (1996). *Science Assessment Framework for the 1996 National Assessment of Educational Progress.*

NAGB (1994). *Science Assessment and Exercise Specifications for the 1994 National Assessment of Educational Progress.*

National Assessment Governing Board (1996). *Policy statement on redesigning the National Assessment of Educational Progress*. Washington, DC: Author.

National Center for Education Statistics (1996). *An operational vision for NAEP - Year 2000 and beyond*. Washington, DC: Author.

Petersen, N. S., Kolen, M. J., & Brennan, R. L. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd Ed., pp. 221-262). New York: Macmillan.

Pommerich, M., & Nicewander, W.A. (1996, June). *Estimating average domain scores*. Paper presented at the Annual Meeting of the Psychometric Society, Banff, Alberta, Canada.

Pommerich, M., & Nicewander, W.A. (1997). *Estimating average domain scores*. Unpublished manuscript.

REFERENCES, CONT.

Rindskopf, D. (1992). A general approach to categorical data analysis with missing data, using generalized linear models with composite links. *Psychometrika, 57*, 29-42.

Rosenbaum, P. R., & Thayer, D. T. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology, 40*, 43-49.

Swinton, S S. (1997). *Block position effects as indirect indicator of motivation in the 1996 NAEP science assessment.* Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, March).

Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item analysis and test scoring using item response theory* (Version 6.1). Mooresville, IN: Scientific Software.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* New York: John Wiley & Sons.

Worthington, L., & Donoghue, J. R. (1997). *Detecting and describing local item dependence using IRT models: Lessons from the 1996 NAEP science assessment.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, March).

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*, 103-118.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40*, 663-705.

Woodruff, D.J., & Hanson, B.A. (1996). *Estimation of item response models using the EM algorithm for finite mixtures* (Research Report No. 96-6). Iowa City, IA: American College Testing.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*, 10-16.

Listing of NCES Working Papers to Date

Please contact Ruth R. Harris at (202) 219-1831
if you are interested in any of the following papers

Number	Title	Contact
94-01 (July)	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94-02 (July)	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk
94-03 (July)	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94-04 (July)	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94-05 (July)	Cost-of-Education Differentials Across the States	William Fowler
94-06 (July)	Six Papers on Teachers from the 1990-91 Schools and Staffing Survey and Other Related Surveys	Dan Kasprzyk
94-07 (Nov.)	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
95-01 (Jan.)	Schools and Staffing Survey: 1994 Papers Presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95-02 (Jan.)	QED Estimates of the 1990-91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95-03 (Jan.)	Schools and Staffing Survey: 1990-91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk
95-04 (Jan.)	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95-05 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
95-06 (Jan.)	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95-07 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95-08 (Feb.)	CCD Adjustment to the 1990-91 SASS: A Comparison of Estimates	Dan Kasprzyk
95-09 (Feb.)	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95-10 (Feb.)	The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95-11 (Mar.)	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95-12 (Mar.)	Rural Education Data User's Guide	Samuel Peng
95-13 (Mar.)	Assessing Students with Disabilities and Limited English Proficiency	James Houser
95-14 (Mar.)	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95-15 (Apr.)	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95-16 (Apr.)	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95-17 (May)	Estimates of Expenditures for Private K-12 Schools	Stephen Broughman
95-18 (Nov.)	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96-01 (Jan.)	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-02 (Feb.)	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk
96-03 (Feb.)	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
96-04 (Feb.)	Census Mapping Project/School District Data Book	Tai Phan
96-05 (Feb.)	Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey	Dan Kasprzyk
96-06 (Mar.)	The Schools and Staffing Survey (SASS) for 1998-99: Design Recommendations to Inform Broad Education Policy	Dan Kasprzyk
96-07 (Mar.)	Should SASS Measure Instructional Processes and Teacher Effectiveness?	Dan Kasprzyk
96-08 (Apr.)	How Accurate are Teacher Judgments of Students' Academic Performance?	Jerry West
96-09 (Apr.)	Making Data Relevant for Policy Discussions: Redesigning the School Administrator Questionnaire for the 1998-99 SASS	Dan Kasprzyk
96-10 (Apr.)	1998-99 Schools and Staffing Survey: Issues Related to Survey Depth	Dan Kasprzyk
96-11 (June)	Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with comments on School Reform, Governance, and Finance	Dan Kasprzyk
96-12 (June)	Predictors of Retention, Transfer, and Attrition of Special and General Education Teachers: Data from the 1989 Teacher Followup Survey	Dan Kasprzyk
96-13 (June)	Estimation of Response Bias in the NHES:95 Adult Education Survey	Steven Kaufman
96-14 (June)	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-15 (June)	Nested Structures: District-Level Data in the Schools and Staffing Survey	Dan Kasprzyk
96-16 (June)	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
96-17 (July)	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
96-18 (Aug.)	Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children	Jerry West
96-19 (Oct.)	Assessment and Analysis of School-Level Expenditures	William Fowler
96-20 (Oct.)	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96-21 (Oct.)	1993 National Household Education Survey (NHES:93) Questionnaires: Screener, School Readiness, and School Safety and Discipline	Kathryn Chandler
96-22 (Oct.)	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
96-23 (Oct.)	Linking Student Data to SASS: Why, When, How	Dan Kasprzyk
96-24 (Oct.)	National Assessments of Teacher Quality	Dan Kasprzyk
96-25 (Oct.)	Measures of Inservice Professional Development: Suggested Items for the 1998-1999 Schools and Staffing Survey	Dan Kasprzyk
96-26 (Nov.)	Improving the Coverage of Private Elementary-Secondary Schools	Steven Kaufman
96-27 (Nov.)	Intersurvey Consistency in NCES Private School Surveys for 1993-94	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

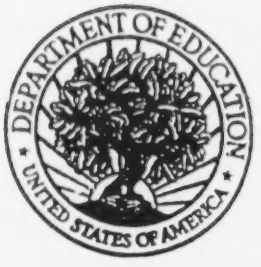
Number	Title	Contact
96-28 (Nov.)	Student Learning, Teaching Quality, and Professional Development: Theoretical Linkages, Current Measurement, and Recommendations for Future Data Collection	Mary Rollefson
96-29 (Nov.)	Undercoverage Bias in Estimates of Characteristics of Adults and 0- to 2-Year-Olds in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
96-30 (Dec.)	Comparison of Estimates from the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-01 (Feb.)	Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association	Dan Kasprzyk
97-02 (Feb.)	Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-03 (Feb.)	1991 and 1995 National Household Education Survey Questionnaires: NHES:91 Screener, NHES:91 Adult Education, NHES:95 Basic Screener, and NHES:95 Adult Education	Kathryn Chandler
97-04 (Feb.)	Design, Data Collection, Monitoring, Interview Administration Time, and Data Editing in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-05 (Feb.)	Unit and Item Response, Weighting, and Imputation Procedures in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-06 (Feb.)	Unit and Item Response, Weighting, and Imputation Procedures in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-07 (Mar.)	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97-08 (Mar.)	Design, Data Collection, Interview Timing, and Data Editing in the 1995 National Household Education Survey	Kathryn Chandler

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
97-09 (Apr.)	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
97-10 (Apr.)	Report of Cognitive Research on the Public and Private School Teacher Questionnaires for the Schools and Staffing Survey 1993-94 School Year	Dan Kasprzyk
97-11 (Apr.)	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97-12 (Apr.)	Measuring School Reform: Recommendations for Future SASS Data Collection	Mary Rollefson
97-13 (Apr.)	Improving Data Quality in NCES: Database-to-Report Process	Susan Ahmed
97-14 (Apr.)	Optimal Choice of Periodicities for the Schools and Staffing Survey: Modeling and Analysis	Steven Kaufman
97-15 (May)	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97-16 (May)	International Education Expenditure Comparability Study: Final Report, Volume I	Shelley Burns
97-17 (May)	International Education Expenditure Comparability Study: Final Report, Volume II, Quantitative Analysis of Expenditure Comparability	Shelley Burns
97-18 (June)	Improving the Mail Return Rates of SASS Surveys: A Review of the Literature	Steven Kaufman
97-19 (June)	National Household Education Survey of 1995: Adult Education Course Coding Manual	Peter Stowe
97-20 (June)	National Household Education Survey of 1995: Adult Education Course Code Merge Files User's Guide	Peter Stowe
97-21 (June)	Statistics for Policymakers or Everything You Wanted to Know About Statistics But Thought You Could Never Understand	Susan Ahmed
97-22 (July)	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
97-23 (July)	Further Cognitive Research on the Schools and Staffing Survey (SASS) Teacher Listing Form	Dan Kasprzyk
97-24 (Aug.)	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97-25 (Aug.)	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
97-26 (Oct.)	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
97-27 (Oct.)	Pilot Test of IPEDS Finance Survey	Peter Stowe
97-28 (Oct.)	Comparison of Estimates in the 1996 National Household Education Survey	Kathryn Chandler
97-29 (Oct.)	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Steven Gorman
97-30 (Oct.)	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Steven Gorman


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA